

DOCUMENT RESUME

ED 241 560

TM 840 076

AUTHOR Donlon, Thomas F.
TITLE An Exploratory Study of the Implications of Test Speededness.
INSTITUTION Educational Testing Service, Princeton, N.J.;
Graduate Record Examinations Board, Princeton, N.J.a,
REPORT NO GREB-76-9P
PUB DATE Mar 80
NOTE 63p.; For related document, see TM 840 089.
AVAILABLE FROM Educational Testing Service, Publications Order
Services, Dept. 1-101, Princeton, NJ 08541.
PUB TYPE Reports - Research/Technical (143)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Estimation (Mathematics); Evaluation; Mathematical
Models; *Measurement Techniques; *Research
Methodology; *Testing; Testing Problems; Testing
Programs; *Timed Tests
IDENTIFIERS Biserial Correlation; *Speededness (Tests)

ABSTRACT

To evaluate test speededness and to derive implications for test program activity, this study reviewed the literature on speed and power, identifying four major approaches to the assessment of speed: the Gulliksen approach, the Cronbach and Warrington approach, the Stafford approach, and the approach of the Educational Testing Service (ETS) as described by Swineford. Two new models were formulated: the Reilly-Donlon approach and the biserial methods. These new methods provide correlations that are analogous to those of Cronbach and Warrington but derivable from a single, conventional administration of a test. The new methods depend centrally on the assumption of normality in the rates of work. Evidence from ETS program files in connection with this hypothesis was evaluated. Methods for estimating the parameters of the distribution of rates of work by graphic methods and the use of normal probability paper were studied. Extensions of these methods to two practical problems were explored: (1) The estimation of time available for review as an ancillary consequence of test speededness and (2) the estimation of the time demands of individual passages and items. The potential implications of omissions for the evaluation of test speededness were demonstrated. (Author/PN)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED241560

TM 840 076

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as
received from the person or organization
originating it.

Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

H. Weidenmiller

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

GRE

GRADUATE RECORD EXAMINATIONS

AN EXPLORATORY STUDY OF
THE IMPLICATIONS OF TEST SPEEDEDNESS

Thomas F. Donlon

GRE Board Professional Report GREB No. 76-9P

March 1980

This report presents the findings of a research
project funded by and carried out under the
auspices of the Graduate Record Examinations
Board.



EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

An Exploratory Study of
The Implications of Test Speededness

Thomas F. Donlon

GRE Board Professional Report GREB No. 76-9P

March 1980

ABSTRACT

Test speededness is a major concern of testing programs from the stand-points of fairness, psychometric efficiency and administrative convenience. The current study was a frankly exploratory effort to systematically evaluate several possible approaches to the appraisal of test speededness and to derive implications for test program activity and for further research.

The study reviewed the literature on speed and power, preparing an annotated bibliography of articles and identifying four major approaches to the assessment of speed. Two new models: The Reilly-Donlon approach and the biserial methods, were also formulated. These new methods have advantages in providing correlations analogous to those of Cronbach and Warrington but derivable from a single, conventional administration of a test.

The new methods depend centrally on the assumption of normality in the rates of work. Evidence from ETS program files in connection with this hypothesis was evaluated. Methods for estimating the parameters of the distribution of rates of work by graphic methods and the use of normal probability paper were studied. Extensions of these methods to two practical problems were explored: 1) The estimation of time available for review as an ancillary consequence of test speededness and 2) the estimation of the time demands of individual passages and items. The potential implications of omissions for the evaluation of test speededness was demonstrated. The implications of the work for program practice and for future research are discussed.

An Exploratory Study of the Implications of Test Speededness

Introduction

The distinction between speed and power in testing has largely followed intuitive lines. There has been general agreement on classical models of purely speeded tests or purely power tests, but little systematic work on the appraisal of tests that are mixtures of the two. The present study was undertaken as a frankly exploratory effort to derive new and useful ways of describing test speededness. In particular, it was focused on a review of problems of speededness as they relate to the evaluation of tests and on suggestions for further research that could contribute to the understanding of these problems.

There are three broad reasons for attending to speed and power: (1) issues of fairness or equity, (2) issues of psychometric efficiency, and (3) issues of administrative efficiency. These three facets of the problem are interrelated but differentiable. A testing program may design its tests, or modify them, with respect to speed and power, in order to achieve goals in each of these three areas.

The first area, the notion of fairness or equity, is a fundamental one. If two candidates work on an examination of 100 items for 40 minutes, and candidate A reads and responds to 80 items while candidate B reads and responds to 40 items, there is a clear potential advantage to A. The test developer cannot overlook this possible advantage. Did B understand the test? Is B as familiar with the testing situation? If B has truly a characteristic slower rate of work that is not easily accelerated, is the resulting distinction between B and A valid, in the sense that B will not do as well on criterion tasks for which the test may be predictive?

This note of validity blends the discussion of equity into the discussion of psychometric efficiency. Speed and power can be established as separate factors or sources of variance in test scores. To the extent that these factors are differentially predictive of a criterion, they may be differentially valuable to the test designer. If a speeded test performance is more predictive of a criterion than a power test performance is, then the test planner will establish conditions, in terms of number of items and time allowed, that foster a speeded performance.

Even if there is no difference between a speeded test and a power test in the prediction of criterion performance, the time efficiency of a speeded test may be of value. That is, it may offer more measurement per minute. In the design of multitest batteries, requiring several hours of testing, where the proper allocation of time to tests is a problem, the ability to elect a speeded test may be a distinct advantage.

Thus, psychometric efficiency shades into administrative aspects. The practical constraints that govern the design of tests are many, and a limit on the available time is a good example. But there are other aspects as well. For one, there is the greater demand of speeded tests for a precise timing. If proctors and administrators cannot meet these somewhat more demanding standards, then a speeded test may be vulnerable to error where a power test is not. Another example of an administrative problem is considered at some length later in this report. It concerns the availability of time for review, or time remaining after answers have been given to all questions. A problem with power tests in test batteries is that everyone must wait to start a subsequent test at the same time. Thus, the fastest workers on the first test will have time remaining, which can cause problems for the proctors.

All three of these issues: equity, psychometrics, and administrative efficiency, have received attention in the measurement literature. Yet out of this considerable body of work there has yet to emerge any widely accepted approach to the description of test speededness. There is no general or common set of indicators that are offered in connection with test evaluation. In particular, there has been relatively little work on the appraisal of speed and power relationships within a single test administration. This lack of a convenient indicator, widely used, impedes the aggregation and exchange of information on test speededness. Furthermore, the indicators suggested or applied by measurement workers are often seemingly limited in their value, leaving uncertainties as to some of the important questions of equity, psychometric utility, or administrative efficiency for the tests they describe.

An analog to this need for a single-administration index may be available in the Kuder-Richardson coefficients. These widely used approaches to test reliability enable the test worker to estimate reliability without the inconvenience of any repeated measurement or parallel form. But the relationship between speed and power, in a correlational sense, has not been describable through a single administration. One major focus of this study has been to consider ways of achieving this and to offer evaluations of their potential.

The study has not involved any significant collection of empirical data. Its purpose was not to discover the specific speed-power characteristics of any tests, or to evaluate any specific test. Rather, it has focused on ways of describing tests in terms of speed and power. What indices have been suggested? How useful are they? What are their limitations? What other indices might be suggested?

There is, of course, a significant body of empirical data concerning tests available at ETS. Each major test administration produces a test analysis, and the raw data for this analysis are of significant value as sources of information about rates of work, etc. The study drew widely upon these program files in its efforts to evaluate certain indices.

There are obvious limits to the usefulness of available data in the testing of hypotheses about speed, power and their description. The study found much useful data available from the existing files, but there are important questions raised by its findings, questions which can only be resolved by further experimental work. The report contains recommendations for such further work.

While centered on the GRE, the study was not limited to this program. It was recognized that there would be data available from other programs at ETS that would be relevant to the topic of speededness. This additional information would be drawn upon, as appropriate, to clarify questions and issues that might not be reflected in the GRE program files. For example, no measure within the GRE program is a purely speeded test. There has been some experience with highly speeded measures in other programs, however. Thus, the scope of the inquiry was broadened to include other relevant information.

There is a sizable literature on the topics of speed and power, reflecting the general importance of the area. A selective bibliography of the major empirical and theoretical studies was developed as a part of this project*. The topic is generally discussed in the major texts on measurement, and data regarding speededness are a regular part of test evaluation. It is appropriate, however, to begin this report with some of this general background.

In sum, this study sought to answer the following questions, essentially in the sequence in which they are presented:

- 1) What major indications of test speededness have been offered in the literature? What are their strengths and weaknesses?
- 2) What other indicators might be suggested?
- 3) What evidence exists for the feasibility, utility, and validity of these new indicators?
- 4) What additional research might be needed to assist in the implementation of these indicators?

Major Approaches to the Evaluation of Test Speededness

The present study identified four major approaches to the description of test speededness in the literature. These are the Gulliksen approach, the Cronbach and Warrington approach, the Stafford approach, and the approach of ETS as described by Swineford. These are briefly reviewed, in turn, in the following sections.

* It is available as GRE Report No. 76-9aP

Gulliksen's approach

A pure speed test is one which measures merely rate of work; a pure power test measures the level of difficulty of the tasks that can be successfully completed, assuming that there is no limit on the time. Gulliksen (1950) has formalized this thinking. He describes two variance ratios:

$$s_u^2/s_x^2 \text{ and } s_w^2/s_x^2, \text{ where } s_u^2 \text{ is}$$

the variance of the number of items unattempted, s_w^2

is the variance of the number of items answered wrongly,

and s_x^2 is the variance of the number of items

answered correctly.

The logic for these ratios is as follows: Assuming a test of N items, scored 1 point for each item answered correctly, then $N = X + W + U$. That is, for each of the N items the answer is either right (X), or wrong (W), or unattempted (U). Since the variance of N is zero, as a constant, the variance of (W + U) taken together, is equal to the variance of X. But the variance of W in a pure speed test is zero, for everyone succeeds on every item attempted. So

$$s_{w+u}^2 = s_w^2 \text{ and the ratio } s_u^2/s_x^2 \text{ is 1.}$$

Conversely, when the variance of the unattempted items approaches zero,

$$s_{w+u}^2 = s_w^2, \text{ and the ratio } s_w^2/s_x^2 \text{ approaches 1.}$$

As s_w^2/s_x^2 approaches unity, we have a power test.

Thus, we have rough indices of test speededness in these ratios.

In practice, however, tests are seldom purely speed or power, and the interpretation of the two ratios is complicated by the effects of correlations among X, W, and U. The values of the ratios can vary considerably. In his book, Gulliksen provides examples of simultaneously large ratios, and describes them as "not very helpful."

Cronbach and Warrington's Approach

The other major approach to a description of test speededness was provided by Cronbach and Warrington (1951). They suggested an index, Tau, defined as follows:

$$\text{Tau} = 1 - \frac{r_{S_A P_B} r_{P_A S_B}}{r_{S_A S_B} r_{P_A P_B}}$$

where the r's are correlations, S and P are speed and power conditions, and A and B are parallel forms of the test. The quantity in the ratio is essentially the squared average correlation between speed and power, corrected for attenuation.

The merit of this approach lies in its evaluation of speededness not in terms of observed magnitude of its variance but in terms of the extent to which it is in agreement with measures secured under power conditions. The index, Tau, approaches zero as the correlations between the speed condition true score and the power condition true score approach 1.0. It essentially asks not "How much speededness is there?" but "How different is the speed measure from the power measure?" It does not, therefore, give us a measure of pure speed or pure power, as Gulliksen's ratios do. A pure speed test, in Gulliksen's sense, subjected to the Cronbach and Warrington procedure, would yield a Tau of 1.0. There could be no speed-power correlation, because the power scores would all be the same, perfect scores on the test. A pure power test, in the sense of Gulliksen, cannot be known in advance to yield any specified level for Tau. It could be that the value of Tau will be lower as respondents under speed conditions fail to win score points that they could win under power conditions. But to the extent that rate of work on items is correlated with the ability to do more difficult items, a high value of Tau could be observed.

The Cronbach and Warrington single index avoids the problems of ambiguity which Gulliksen's ratios have for a broad middle ground of tests. However, the method requires two administrations and two parallel forms in order to secure the four needed correlations. These requirements have limited its usefulness. An approximation to the Cronbach and Warrington index is sometimes obtained by the simple device of administering a test under a time limit and then allowing additional time. With the additional time, the test becomes a power test. Some way of distinguishing marks made under the two conditions is needed. Different colored pencils, different answer sheets, or a line following the last speed-condition entry have been used. This approximation technique has the difficulty that the errors of measurement between speed and power are correlated. Furthermore, while it requires less special administrative activity than the complete Cronbach and Warrington procedure, it is still a more complicated procedure than the routine administration.

Stafford's Approach

The Stafford approach is to develop a "speededness quotient," as follows:

$$SQ = \frac{U}{W + U}$$

where U = the number of Unattempted items
W = the number of Wrong items

The simplicity of this approach is apparent, and is a major virtue of the method. In dealing with "pure" tests, the results will be consistent with Gulliksen's. When a test is a pure speed test, there are no wrong answers; W = 0, and $SQ = U/U = 1.0$. When a test is a pure power test, there are no unattempted items; U = 0 and $SQ = 0/W = 0.0$.

Stafford (1971) suggests that his methods may be applied to individuals as well as groups. As will be seen, something like the Stafford index is implicit in one aspect of the ETS approach to speed, described below. It is plausible that a test performance with no wrong answers is fundamentally reflective of speed. If two individuals have the same high SQ, but different scores, the difference in score seems most plausibly attributable to speed.

But the greatest weakness in the Stafford approach seems to be its failure to describe the variation among people. In terms of the concerns for test fairness, or for psychometric efficiency, it is the amount of variance attributable to speed or power, and the potential relation of this variance to criterion performance, that is meaningful. The Stafford methods, centering on measures of central tendency, offer little information on variance.

The ETS Approach

The ETS approach to speed is essentially a judgment based upon certain more or less arbitrary rules of thumb. Table 1 presents a typical description of speededness from an ETS test analysis report. As it indicates, three indicators are focused upon: (1) the percentage completing the test, (2) the percentage of the group completing 75% of the test, and (3) the number of items completed by 80% of the candidates. These three values are used to appraise speed. As Swineford (1974) indicates:

...It has long been the policy at ETS to regard a test as essentially unspeeded if at least 80 percent of the group reach the last item and if virtually every one reaches at least three-quarters of the items. These are arbitrary criteria and should not, of course, be too strictly applied. They are useful from a practical point of view, because the 80 percent who finish is likely to include all the able students; if additional time were allowed, the other 20 percent would be unlikely to increase their scores, particularly when the usual practice of placing difficult items at the end of the test is followed. Moreover, the test supervisors would be faced with the problem of a restless group, eager to get away from the examination room. It is important to understand, however, that if the criteria are not met it does not necessarily follow that the test is speeded...

The criteria are essentially signals of possible or potential speededness. When the results conform to the criteria, no major problem of speededness is likely; most candidates complete most items, and the possibility of a dramatic change in the relative standing of a candidate, given additional time, is not likely.

TABLE 1

ETS REPORTS ON SPEEDEDNESS: BASIC DATA

(GRE Psychology Form XGR)

Speededness of Test

Percentage completing test	58.0
Percentage completing 75 % of test	99.5
Number of items reached by 80% of the candidates	199.0
Total number of items	200.0

The ETS analysis recognized the need to supplement these basic data, and offers for each test a summary page such as the one in Table 2. In reviewing the application of such information, the ETS Technical Manual provides the following discussion:

The question of possible speededness can be investigated further in connection with the special score data.... There are frequency distributions of the number of items answered right, answered wrong, omitted, and not reached and a bivariate distribution of Score versus $R + W$. (The term "two-way" is used instead of "bivariate" in the reports for the sake of readers who may not be familiar with many mathematical terms.) It is customary to add to the two-way table two or three lines. First, the principal diagonal passes through equal values of Score and $R + W$. If speed is an important element in the score, there will be a noticeable tendency for the entries in the table to lie close to the principal diagonal. A substantial blank area between the line and the configuration of entries indicates a power measure, as does even more a strong tendency for the entries to lie in the right-hand columns. Secondly, a dashed line below which 99 percent of scores that are obtained by chance marking of the answer sheet would fall is drawn. This line is a curve, but the curvature is so slight for the higher values of $R + W$ that it can be drawn with a ruler. Tables for help in locating the line are appended. It should not be assumed that scores below the line are all chance scores. The majority or all are most probably scores earned by candidates for whom the test is too difficult to provide a dependable score. A large proportion of scores within the chance area defined by the dashed line shows the test to be too difficult for the group as a whole, unless interest is restricted to high-scoring candidates. A third, solid line is drawn when $\text{Score} = R$. It passes through the most probable chance score for successive values of $R + W$. When $\text{Score} = R - KW$, the most probable score is usually zero. When a test contains items that vary with respect to the number of alternatives, only the main diagonal is shown. The two chance lines can be indicated only for the total number of items.

Evaluating Descriptions of Speed and Power

Both the approaches of Gulliksen and of Cronbach and Warrington are appraisals of the overt impact of speed on test scores. The Gulliksen approach essentially tests the magnitude of the variance of the unattempted items; the Cronbach and Warrington approach tests the correlation between "scores with a time limit" and scores without one. Both approaches are centered on a specific configuration of test material and allowed time. The ETS approach is more complex, but primarily follows the Gulliksen model, in that no correlation of speed and power is formally estimated.

TABLE 2

ETS REPORTS ON SPEEDEDNESS: SUPPLEMENTARY DATA
(GRE Advanced Chemistry Form 3BCR)

R - W SCORE	34 TO 40	41 TO 47	48 TO 54	55 TO 61	62 TO 68	69 TO 75	76 TO 82	83 TO 89	90 TO 96	97 TO 103	104 TO 110	111 TO 117	118 TO 124	125 TO 131	132 TO 138	139 TO 145	146 TO 150	SCORE f	R (RIGHT) Number of Items f		W (WHONG) Number of Items f		O (OMIT) Number of Items f		N R (NOT REACHED) Number of Items f		
	40	47	54	61	68	75	82	89	96	103	110	117	124	131	138	145	150	f		f		f		f		f	
119-124																1		1123-129	1							86- 64	2
111-116															1	2		3116-122	4	126-133		1	105-111	1	75- 79		
103-110															3	2	2	7109-115	13	118-125		1	99-104	3	70- 74		
95-102												2		3	9	6	3	25102-108	27	110-117		4	91- 97	6	65- 69		
87- 94											1	1	6	7	9	7	6	37	95-101	36	102-109		10	84- 90	6	60- 64	4
79- 86											7	3	9	10	7	9	10	55	88- 94	56	94-101		14	77- 83	20	55- 59	1
71- 78										2	6	14	19	13	24	7	8	90	81- 87	87	86- 93		33	70- 76	25	50- 54	3
63- 70							2	2	2	6	8	11	11	17	10	8	2	79	74- 80	95	78- 85		40	63- 69	60	45- 49	4
55- 62						1	1	7	3	12	17	24	22	27	20	9	3	146	67- 73	130	70- 77		71	56- 62	64	40- 44	6
47- 54						2	4	10	9	13	8	22	27	13	16	2	6	132	60- 66	130	62- 69		139	49- 55	78	35- 39	9
39- 46			1	1	6	1	9	17	24	15	20	19	16	17	10	4	1	161	53- 59	140	54- 61		172	42- 48	115	30- 34	4
31- 38				2	4	5	14	10	19	19	16	14	14	13	8	7	4	151	46- 52	142	46- 53		220	35- 41	129	25- 29	15
23- 30	1		1		3	2	6	14	12	14	15	10	6	10	2	2	4	102	39- 45	127	38- 45		185	29- 34	125	20- 24	19
15- 22		4	4	5	5	5	8	6	7	13	12	4	3	11	5	1	2	95	32- 38	85	30- 37		158	21- 27	172	15- 19	31
7- 14	1	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	50	25- 31	55	22- 29		72	14- 20	144	10- 14	45
-1- 6		1			3	1	1	3	1	1		3	2	4	1		2	23	18- 24	25	14- 21		33	7- 13	120	5- 9	55
-9- -2												1				1	1	3	11- 17	7	6- 13		7	0- 6	86	0- 4	962
TOTAL	2	7	7	9	24	19	52	74	81	96	116	131	138	146	130	70	58	1160	1160		1160		1160		1160		
SCORE	R-.250CW																	MEAN	48.16	60.97	51.72	33.59	3.73				
																		STANDARD DEVIATION	23.30	21.20	18.58	20.88	9.17				

The Cronbach and Warrington correlational concepts, however, would seem to be fundamental in that when speed and power are highly correlated, the use of either, or their combination is virtually optional with the tester. It is not the magnitude of unattempt variance alone that matters in concerns about speed. It is the extent to which a speeded performance is something different from, something inconsistent with, a power performance.

The Gulliksen approach basically asks is there any evidence of being affected by speed? The Cronbach and Warrington approach basically asks whether, given that some records are affected by speed, there is evidence that things would be different, in the sense of a ranking of candidates, under conditions of power. Of these two aspects, the Cronbach and Warrington concern is the one most frequently overlooked in practice. This is largely due to the difficulty of obtaining an adequate estimate. In a sense, the Cronbach and Warrington question and the Gulliksen question are two different but equally important aspects of the concern for speed.

New Approaches: The Reilly and Donlon Model

A fundamental impetus to the present study was an interest in the potential of a single administration approach to a simple estimate of a speed-power correlation coefficient. In this context, particular attention was devoted to the approach of Reilly and Donlon. Essentially, this approach explores the possibility of generalizing from the observable correlation between speed and power, for those who fail to complete the test, to the correlation that would be observed for the total group. The gist of the method is the assumption of an underlying bivariate normal distribution for a rate of work variable and for the observed score.

The Reilly-Donlon method assumes that each individual has a characteristic rate of work on the test, and that these rates of work are normally distributed. Figure 1 shows the general model.

The assumption of normally distributed test scores is a familiar one. Assumptions about the rate of work of people on tests are not so common. In fact, the distribution of rates of work is seldom formally considered. Attention generally centers on two broad classes of rates: those that enable people to finish, and those that force them to leave some Not Reached responses. The characteristics of the full distribution are held to be more or less unnecessary. The Reilly-Donlon method, on the other hand, considers this full distribution to be fundamental to the analysis.

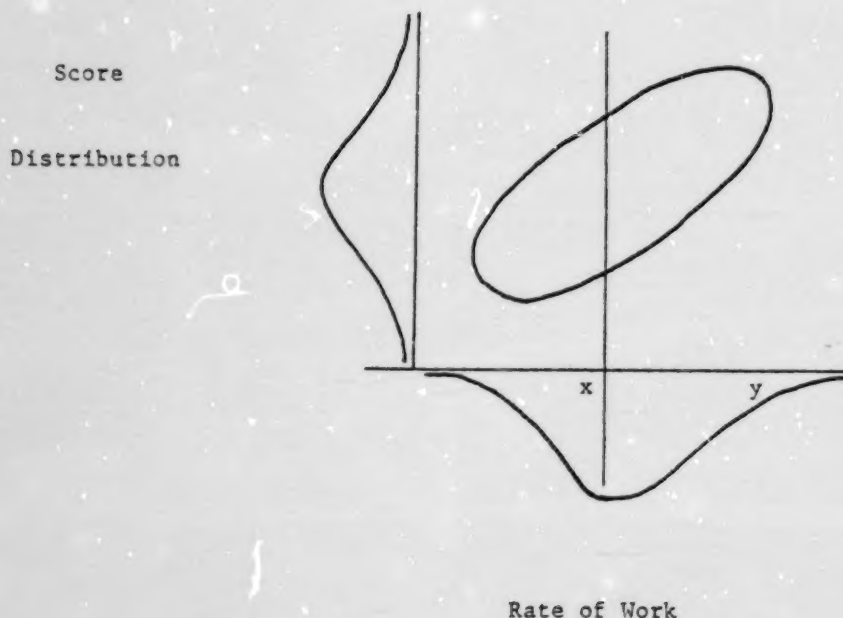


Figure 1. The bivariate normal distribution of score and rate of work.

Figure 1 shows two specific rates of work: x , a rate typical of those dictated by test time limits, and y , a much more demanding rate. A work rate of x items/minute or more enables a person to finish, and only the subgroup of persons who work at rates less than x will yield observable variation in performance. If the time limit were more severe, so that only those persons with a work rate of y would finish, then virtually the entire bivariate distribution of score and rate would be observed. If y were the critical rate, then the full range of values for both variables would be observable, and the correlation could be computed by ordinary methods. But where the observable variation is constrained to a portion of the group, as in the case of a work rate of x , the observed data must be generalized. This is done by applying conventional selection theory formulas for inferring full-group correlations from data for subgroups explicitly selected on one of the variables. In this case, the imposition of a time limit constitutes explicit selection on the rate of work variable.

The normally distributed rates of work produce a normally distributed Not Reached distribution. The concept of a normally distributed Not Reached distribution is made somewhat difficult to conceptualize because the typical test shows only a limited range of Not Reached values, and because all persons finishing the test receive a Not Reached score of zero. But this constancy of score does not mean that they all finished at the same time. In fact, there is a wide variation in work rates. It is this variability that is assumed to be normally distributed. It is sometimes helpful to think of the variable as "items completable" in the allowed time. Persons with Not Reached scores other than zero have "items completable" scores less than the number of items in the test.

The concepts are perhaps made more accessible by conceiving of differences in rates of work as items/minute. The usual NR score is readily translatable to this other variable as the example below shows.

Assuming a test of 30 items, given with a time limit of 30 minutes, the following computations may be made:

Items Completed	Not Reached Score	Items/Minute
30	0	1.000 greater
29	1	0.967
28	2	0.933
27	3	0.900
26	4	0.867
25	5	0.833
24	6	0.800
	etc.	

Thus, the NR score has a linear relationship to the items/minute score, as follows:

$$\begin{aligned} \text{Items/minute} &= 1.000 - (\text{NR})(0.03333) \\ \text{NR} &\geq 1. \end{aligned}$$

NR is accordingly, a measure of rate of work, of "items completable."

Estimating the Mean and Standard Deviation of Items Completable

Given the assumptions that items completable and total score have a bivariate normal distribution and that examinee working speed would not be affected by any additional set of parallel items, to obtain the estimate s_c the following relationships can be employed:

$$(1) \quad p_n = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_c} e^{-z^2/2} dz$$

$$(2) \quad \mu'_1 = \frac{1}{p_n \sqrt{2\pi}} \int_{-\infty}^{z_c} z e^{-z^2/2} dz$$

$$(3) \quad \mu'_2 = \frac{1}{p_n \sqrt{2\pi}} \int_{-\infty}^{z_c} z^2 e^{-z^2/2} dz$$

$$(4) \quad \sigma' = (\mu'_2 - (\mu'_1)^2)^{1/2}$$

$$(5) \quad s_c = s'_c / \sigma'$$

p_n = is the proportion of examinees not finishing the test.

z_c = is the unit normal deviated corresponding to p_n .

μ'_1 = is the first moment of the truncated unit normal distribution with terminus z_c .

μ'_2 = is the second moment of the truncated unit normal distribution.

σ' = is the standard deviation of the truncated normal distribution.

s'_c = is the observed standard deviation of items completed for those not finishing the test.

s_c = is the estimate of the standard deviation of items completed for the complete theoretical distribution (i.e., finishers and nonfinishers with an infinite number of additional parallel items).

Once s_c has been obtained, it becomes possible to obtain M_c , the estimated mean number of items completed, in the equation:

$$(6) \quad M_c = M'_c - s_c \mu'_1$$

where M'_c is the observed mean number of items completed for examinees not finishing the test.

Estimating the Correlation Between Items Completed and Total Score

Given s_c , the following formula can be used to obtain r_{xc} .

$$(7) \quad r_{xc} = \frac{s_c r'_{xc}}{\sqrt{s_c^2 (r'_{xc})^2 + (s'_c)^2 (1 - (r'_{xc})^2)}}$$

where r'_{xc} = the correlation between items completed and total score for nonfinishers.

The foregoing discussion presents the essential formulas for the method. The proportion not finishing the test has observable means and standard deviations for both variables: rate of work in the form of Not Reached and also the test score. The observed Not Reached mean and standard deviation are set equal to the mean and standard deviation of the normal deviation scores, readily derivable from the assumption of a truncated normal distribution. Once the standard deviation for the full distribution of items completed is established, it is used to generalize the observed correlation to an estimate of the correlation in the full group. It is these distribution parameters and correlations that are the essential goal of the method.

To assist in the calculation of the formulas, values of the mean and standard deviation of the truncated normal curve have been computed. They are given in Appendix A. These values are developed for increments of .01 in the proportion of area under the normal curve. Using these tables, it is simple to estimate the mean and standard deviation of the work rates for the total group.

The fundamental test of these formulas must lie in a comparison of the correlations they estimate and empirically gathered correlations. This empirical work was beyond the scope of this study. An effort was made to gather the data needed for calculating the coefficients for a number of ETS tests. Unfortunately, the value of r_{xc} for the group which fails to complete the test cannot be secured without modification of the regular test analysis computer programs. There is currently an interscore correlation matrix including Rights, Wrongs, Omits and Not Reached, but it includes the entire sample, not just those who fail to complete.

The inability to compute sample Reilly-Donlon correlations using existing ETS data was a serious limitation on the present work. The range of values obtained, and the shifts in values across types of tests would be of considerable interest. Nonetheless, the most fundamental question concerning the estimation technique lies in its validity: to what extent do the estimates reflect the values which would be obtained empirically?

Biserial Estimation of the Speed-Power Correlation

The Reilly-Donlon estimates derive from the assumptions of bivariate normality and generalize the covariation within the Not Reached group to infer the correlation for the group as a whole. Another approach to the same basic model of bivariate normality can be based on the fundamental principles of the biserial correlation coefficient.

The required formula is

$$r_{bis_{S-P}} = \frac{m_c - m_t}{s_t} \cdot \frac{p}{y}$$

- where $r_{bis_{S-P}}$ = the biserial estimate of the speed-power correlation.
- M_c = the mean test score of those completing.
- M_t = the mean test score of the total group
- s_t = the standard deviation of the test score of the total group
- p = the proportion who complete the test.
- y = the ordinate of the normal curve above which p percent of the area lies.

An example of this formula may be applied to VGR1 Verbal. Fifty-five percent of the group completed the test. Their mean score was 14.2. The mean score for the total group is 13.0, and the standard deviation of the scores is 4.0. These score values are normalized item analysis criterion measures used in place of raw test scores, but are extremely highly correlated with them.

Applying the formula yields the estimate

$$\begin{aligned} r_{bis_{S-P}} &= \frac{14.2 - 13.0}{4} \times \frac{.55}{.40} \\ &= .41 \end{aligned}$$

As is the case with the Reilly-Donlon estimates, the true value of these coefficients can only be developed through comparisons with empirical estimates. The logic of such a coefficient is such that when the values are sufficiently great it is evidence that the presence of moderate speed variance may be safely disregarded: the same individuals will perform well under either condition.

Appraising the Assumptions of Normality

The new approaches to a single-administration estimate of the speed-power correlation depend greatly on the usefulness of the assumption of normality in the rates of work. The current study could not collect data on this point. Efforts were made to evaluate the hypothesis using program data.

The first approach was a more rigorous one, applying the Chi-square test of agreement between observed and theoretical frequencies for the "items completed" variable. Table 3 presents these data for three mathematics tests used in the College Board Cooperative Guidance and Placement program. The choice of categories for the "items completed" variable is an optional one, and decisions were made for convenience and to avoid empty cells. The results of the approach are consistent: the distributions fail to conform to the hypothesis.

It was recognized, however, that a Chi-square test with large samples is an extremely sensitive test. It was felt, therefore, that a more practical approach would consist of the use of normal probability paper and the appraisal of linearity in the plots.

An example of this approach is provided in Figure 2. These data describe the percentage reaching each item for a special pretest of 65 sentence completion items administered in 30 minutes in the context of the College Board Scholastic Aptitude Test. They show that 99.8% of the group were still working at item 23, and that this percentage declined with consistency so that 21% of the group completed the test. There are minor irregularities in these data, but the general trend seems well described by a linear fit that confirms a reasonable approximation to the normal distribution.

Additional examples of such plots are provided in Figures 3 and 4, which present data for GRE Aptitude subtests, Form XGR. Figure 3 describes the Vocabulary subtest, Figure 4 the Quantitative subtest. The reading test plots exhibit special characteristics, and are discussed below.

The GRE plots in Figures 3 and 4 are based on somewhat more limited data, in the sense of percentage of the group, than is the case in Figure 2. The time limit for the experimental pretest described in Figure 2 was specifically established to induce speededness, so that only 20% of the group completed the test. The time allotments for the GRE tests are much more generous, so that 70% or more of the groups finish working.

TABLE 3
STATISTICAL TESTS OF NORMALITY

Test C1*		
Items Completed	Theoretical Frequency	Observed Frequency
29-33	369.75	254
26-28	416.50	374
22-25	326.00	372
19-24	200.00	272
15-18	95.70	151
0-14	49.60	33

Chi-square 110.49**

Test B1*		
29-34	307.90	301
25-28	253.60	169
0-24	36.70	82

Chi-square 84.29**

Test C2*		
21-24	416.50	397
18-20	326.00	403
15-17	200.00	244
11-14	95.70	90
0-11	49.60	22

Chi-Square 44.50**

*The Tests are Mathematics Tests included in The College Board Cooperative Placement and Guidance Program.

**p<.01.

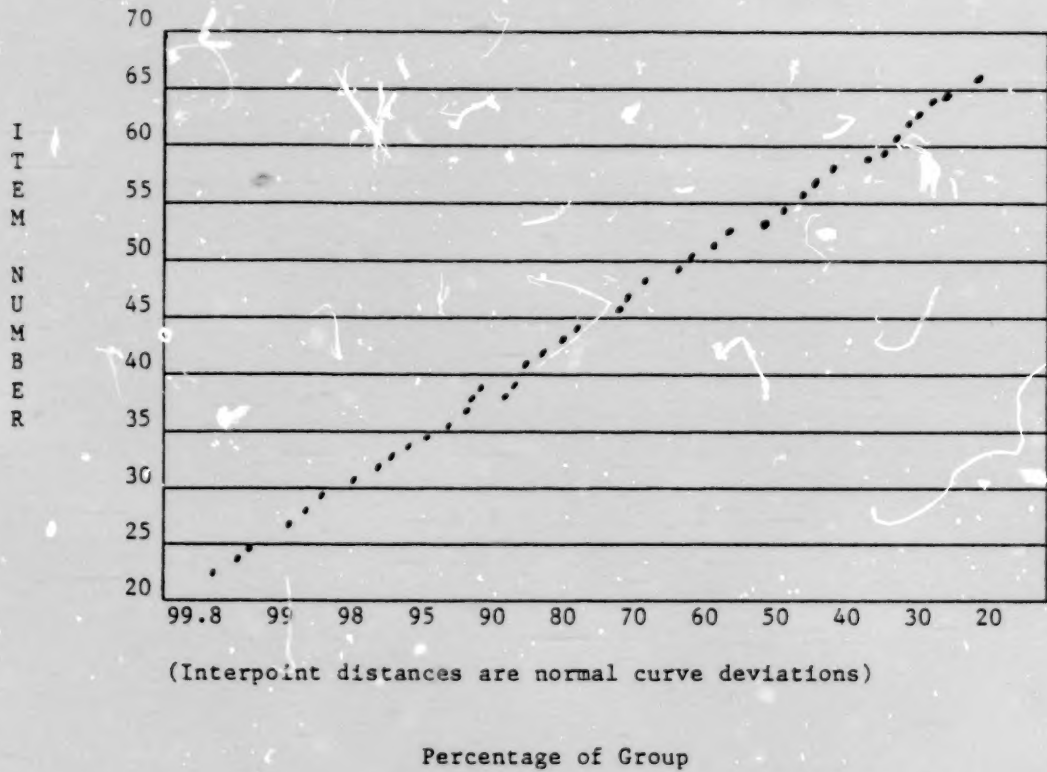


Figure 2. Percentage of group still working at various points in the test - SAT Pretest RSA4S₁₇ (sentence completion items.)

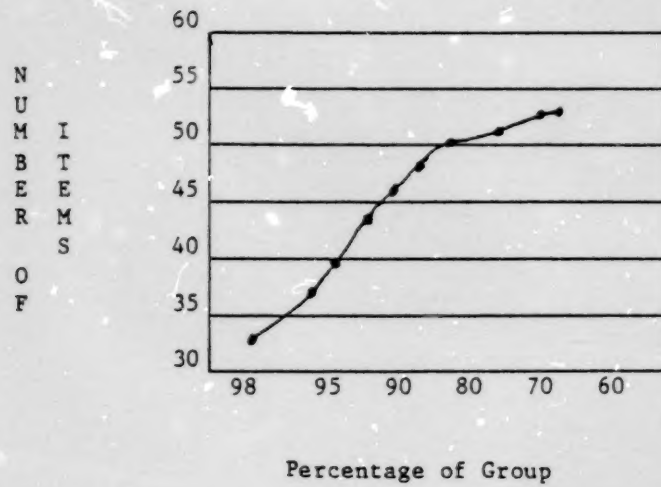


Figure 3. Percentage of group still working - XGR Aptitude Vocabulary subtest.

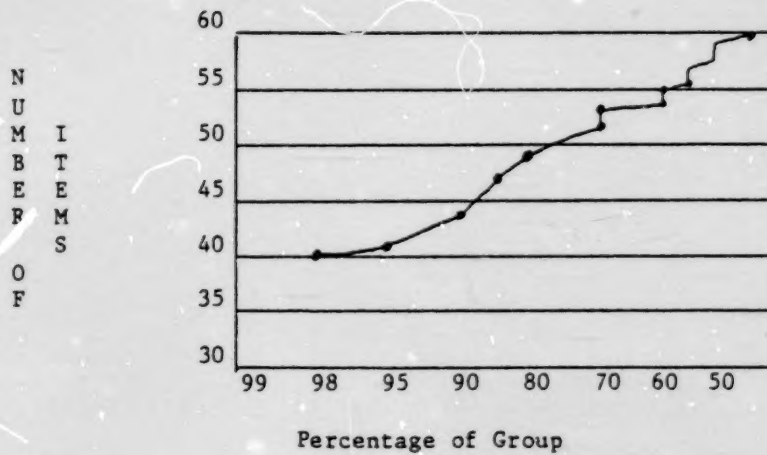


Figure 4. Percentage of group still working - XGR Aptitude Quantitative subtest.

The essence of the single-administration methods is the determination of a live relating item number and percent still working and its extrapolation to the entire group via the assumption of normally distributed rates of work even for those who finish the test. Since many tests show only the 30% or so affected by dropout that is found in Figures 3 and 4, a special effort was made to find a highly speeded measure which would reveal work rates for an entire group. Such a measure was located in the Mosaic Comparisons Test of the College Board Cooperative Guidance and Placement battery. This test is a measure of perceptual speed and accuracy. Each item consists of two squares divided into columns of smaller squares. The object is to locate the column continuing the one square that is shaded differently. An example of the item type is as follows:

DIRECTIONS: This tests consists of pairs of mosaics, that is, patterns of squares like those found on tiled floors or walls. Each mosaic is made up of a number of partially shaded squares. The mosaics in each pair are identical except for one square which differs in shading. The vertical columns of both mosaics are labeled A to C, A to D, or A to E according to the number of columns in the mosaic. Your task will be to locate, for each pair of mosaics, the column that contains the single square which is shaded differently. Then mark the space on your answer sheet that corresponds to the letter at the head of that column.

Sample question

1.

ABC



ABC



Sample answer

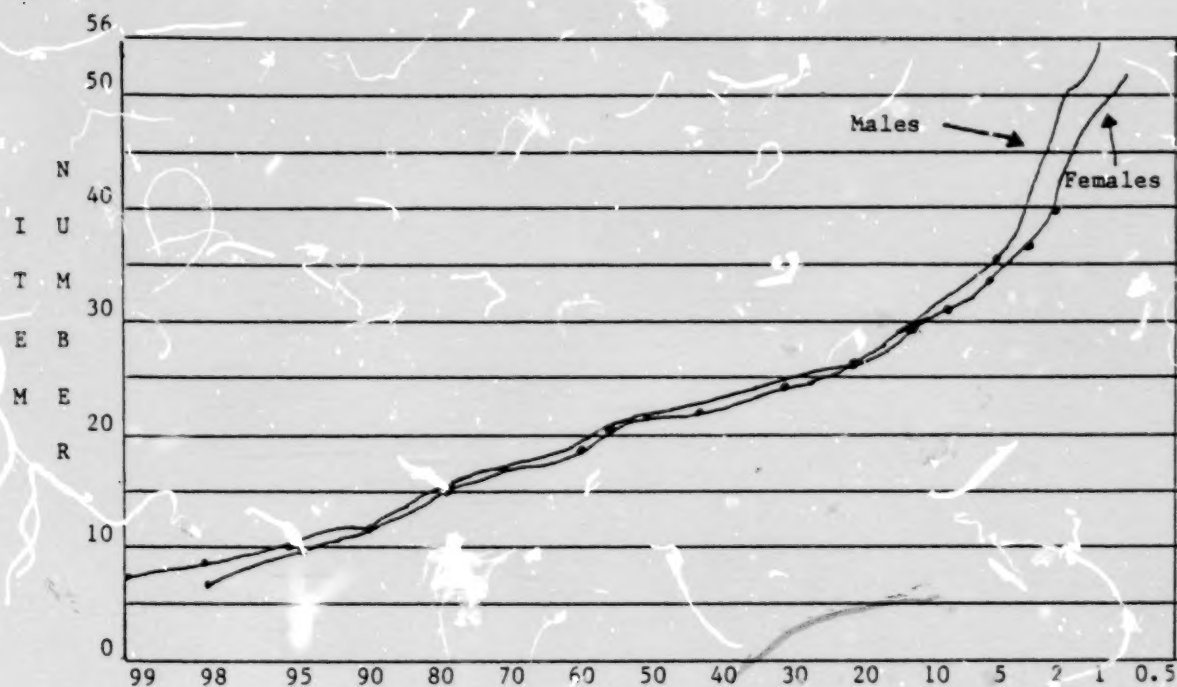
1.

Ⓐ ● Ⓒ

The test is virtually uncompletable within the time limit. Section I of the test, for example, has 56 items and provides a time limit of 3 minutes. This is an average time allowance of 3.2 seconds per item. Figure 5 is a plot of the dropout for this test, separately for males and females. These plots confirm the general failure to finish. Only very small groups reach the last item—about 1.5 percent for males, about 0.5 percent for females.

The principal implication of this graph is its essential support for normality in rates of work. There is a reasonably linear character throughout the major part of the plot, so that 95% of the population seems well described by a normal distribution. Given this evidence, it cannot be concluded with finality that normality describes the unobserved work rates of people who complete a test, but the assumption is supported.

The sharp upward slope towards the end indicates that for both sexes there is a sudden upsurge in rate of work. The top 5 percent of each sex distribution is involved in this pattern. It may indicate an instance



(Interpoint distances are normal curve deviations)

Percentage of Group

Figure 5. Percentage of male and female groups still working - CGP Mosaic Comparison Test.

of blind guessing towards the end of the test. It affects the dropout of the last twenty items or so, and seems to have a kind of "acceleration," an indication of faster and faster work as the time limit approaches. It is interesting to note the contrast between this essentially accelerated performance and the "deceleration," a flattening of the curve, which characterizes the plots of tests with more difficult, time-demanding items and more generous time limits.

Graphic estimates of rate-of-work parameters

The graphic estimates were introduced in the service of testing normality. They also facilitate an informal estimation of the mean and standard deviation of the "items completable" distribution. The mean of the distribution is taken to be that "items completable" score, which is the 50th percentile for the group. The standard deviation of the distribution is taken to be the "items completable" interval between the mean and the 16th (or 84th) percentile for the group.

An example of the method is provided in Figure 6 for the sentence completion data of Figure 2. The line has been drawn in by hand. It intersects the vertical for the 50th percentile roughly at an items completed point of 84. The line intersects the vertical for the 84th percentile roughly at a score of 42. The difference between these two values is 12 points. If the distribution were truly normal, the methods could be used with great precision, for the score points associated with percentiles would be highly accurate. The present data can afford only an approximate estimate. But they will often be convenient for practical work, particularly when an item analysis provides information on dropout for items in sequences.

Figure 7 is another example of the method, this time for XGR Quantitative Aptitude. Here the judgment of the test specialist must be used to evaluate the obvious nonlinearity of the plot. In this case, the last item was not attempted by many people, a decline in percent still working which was not predicted by the preceding items. This item was disregarded in the plotting. The points at the left-most end of the plot are based on very few cases, and similarly have been given less weight in the placement of the line.

The ultimate utility of these methods, their consistence with the formal calculations of the Reilly-Donlon equations, can only be evaluated through empirical studies. Both the formal methods and the graphic methods derive from assumptions of normality. They will offer useful information to the extent that these assumptions are met. The values of the parameters they estimate are in a sense, an independent set of questions. Assuming that normality is justified, and reasonably approximated, assuming that the mean and standard deviation of the items completable distribution is reasonably well estimated, what value do these have for test development and use? The next sections of this report deal with two potential applications.

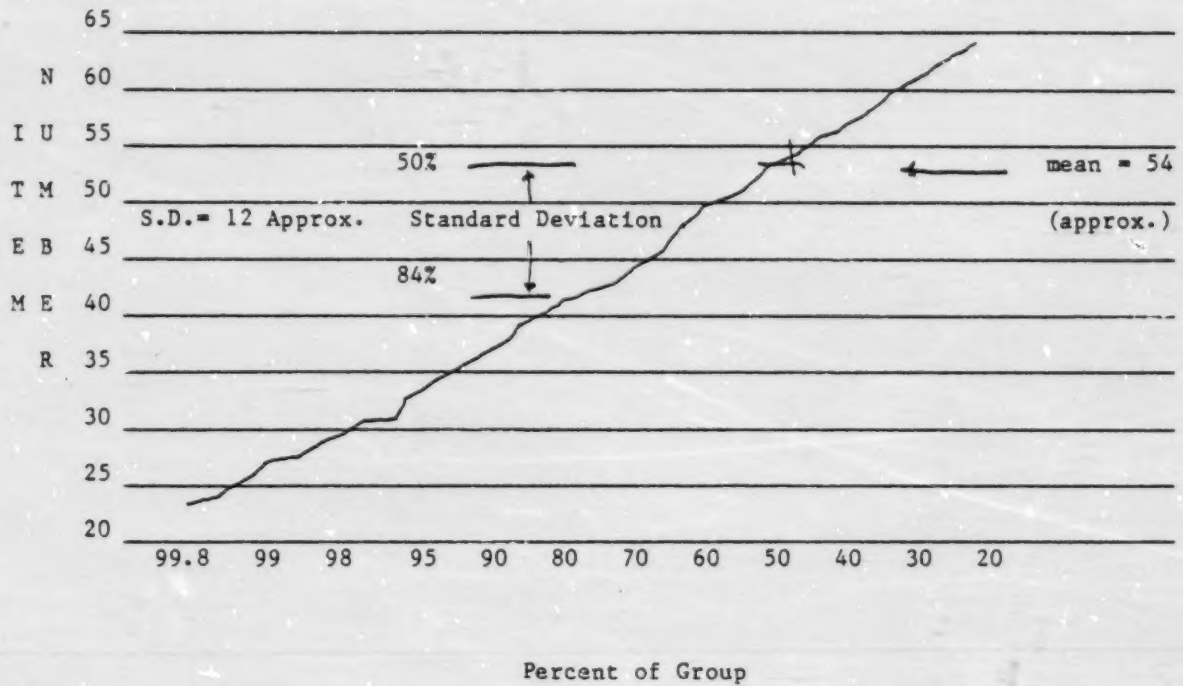


Figure 6. Estimation of the mean and standard deviation by graphic methods - RSAYS₅ Sentence Completion.

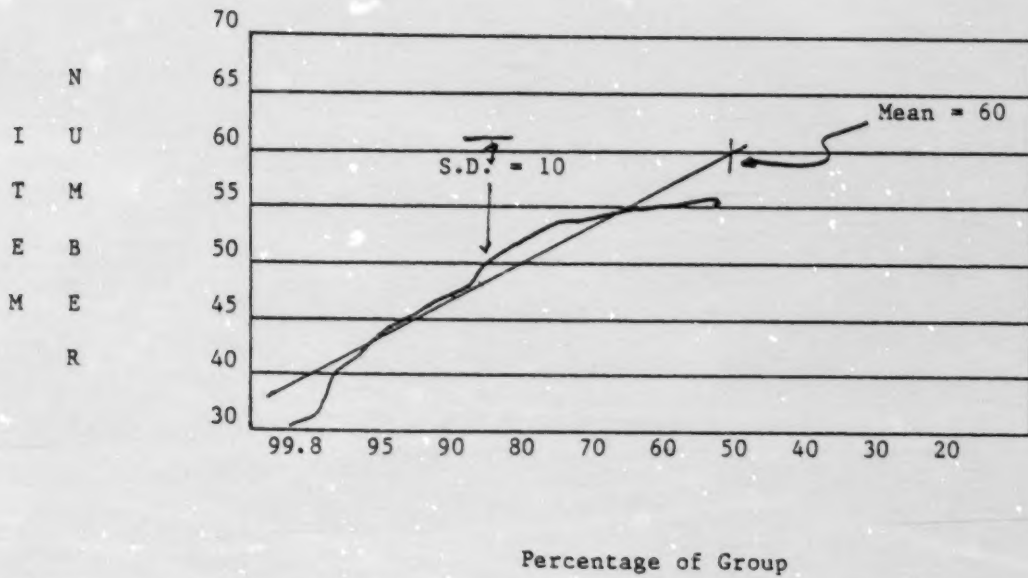


Figure 7. Estimation of the mean and standard deviation by graphic methods - XGR Quantitative Aptitude.

Time for review

The approach to speededness that builds on an assumed normal distribution of work rates has a valuable by-product. To date, most comment on test speededness has centered on the problem of failure to finish. In this context, the basic problem is the usefulness and the fairness of the resulting score distribution, which blends the attainments of finishers and nonfinishers. Are distinctions among individuals which result from rate of work, as evidenced by a failure to finish, valid and meaningful? Can these distinctions be combined with distinctions which are based not on rate of work but on power, on success on items?

Such questions need not be restricted to a consideration of those finishing or not finishing. As the assumption of normality makes clear, there is a wide range of rates of work within the two groups defined by completion. These differences affect the entire test experience. In particular, they lead directly to differences among candidates in the time available for review.

The time available for review is the time remaining, under the official time allotment for the test, subsequent to the candidate's finishing the test. Little is known about this variable as a factor in test performance. The emphasis in speed analysis on the failure to complete has largely obscured the question of time available for review. Furthermore, there has not been a direct indication of this time derivable through an item or test analysis in a manner analogous to the use of the Not Reached variable as a measure of failure to complete. The assumption of normality, however, makes it possible to estimate times for review, for the assumption makes it possible to estimate the times for completion.

Time for review is a direct function of work rate, for time for review is the difference between the time allowed for the examination and the time required to complete it:

$$T_R = T_A - T_C$$

$$T_R = \text{time for review}$$

$$T_A = \text{time allowed}$$

$$T_C = \text{time required to complete}$$

These time values may be conveniently expressed in a number of ways as items/minute, items/allowed time, etc. T_C is determined by the equation.

$$T_C = \frac{n}{r} \text{ where}$$

n = the number of items in the test

r = the rate

The distribution of the time to review will not ordinarily be normal, given the assumption of a normally distributed rate of work, for the T_C values are not a linear function of the work rates. Consider the situation projected in Figure 6. The group confronted a 65-item test. On the basis of the estimates of a mean of 54 items completable and a standard deviation of 12 items, about 18% of the group would be expected to finish the test (about 21% actually did). But when did the earliest finishers complete the test? On the assumption of normality, about 1% of the sample have work rates 2.33 standard deviation units above the mean or greater. Such work rates in this instance would be $54 + (2.33)(12) = 82$ items in 30 minutes, or 2.73 items/minute. Solving for T_C yields $65/2.6 = 23.8$ minutes. Thus, the earliest finishers on this test had about 6.2 minutes or more for review.

The general logic can be used to work out a distribution of time for review. Five minutes for review requires that the test be completed in 25 minutes. This is a rate of $65/25 = 2.6$ items/minute, or, in terms of the "items completable" calculation, 78 items in the allowed time. This is 2.0 standard deviation units above the mean. About 2.3% of the group will show rates as great or greater than this. Four minutes for review requires completion in 26 minutes, a rate of 2.5 items/minute. About 4% of the group show rates as great or greater. These calculations can be continued, and a table developed, as in Table 4.

TABLE 4

Estimated time for review RSA45 ₁₇ - SAT Sentence Completion Pretest	
Time for review	% in category
6.0 min. or more	1.2%
5.0 - 5.99	1.1%
4.0 - 4.99	1.7%
3.0 - 3.99	2.0%
2.0 - 2.99	4.0%
1.0 - 1.99	4.0%
0.0 - 0.99	4.0%

These data give an indication of the distribution of time remaining for review. In this case, since most of the group did not finish the test, the percentages enjoying significant times for review are not great. Figure 8 presents data on a test that was essentially completed by the group working on it. When the methods are applied to the data of Figure 8, the distribution presented in Table 5 is determined.

In Table 5, the percentages of the group enjoying significant time for review are much greater. About 19% of the group may be estimated to have 5 minutes or more for review. The earliest finishers will have 8 minutes or more to go over the test.

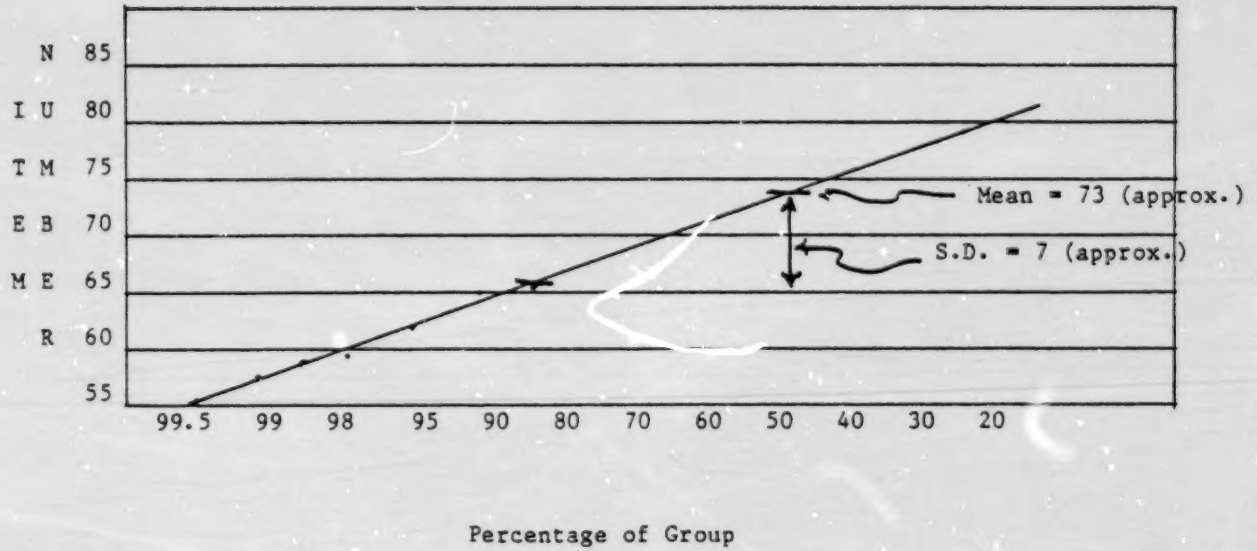


Figure 8. Estimation of the mean and standard deviation
RSA45₂₄ - verbal pretest.

TABLE 5

ESTIMATED TIME FOR REVIEW
RSA45₂₄ - VERBAL PRETEST

Time for Review (in minutes)	% of group in category
8.0 or more	1
7.0 - 7.99	2
6.0 - 6.99	6
5.0 - 5.99	10
4.0 - 4.99	13
3.0 - 3.99	16
2.0 - 2.99	15

These estimates are only approximate. They suggest, however, that time for review is an important concomitant of the variability in work rate. Three aspects of time for review would seem to be important.

First, can candidates convert time for review into increased scores? The evidence on this is not great. Candidate behavior during examinations has not been thoroughly researched. Armchair theorizing has considered some aspects of the situation, with some authorities arguing that candidates should never go back and reconsider an answer, on the ground that these longer deliberations expose the candidate to the risk of succumbing to the seductive misleads of the test developer. More recent research evidence tends to suggest that changes may be sufficiently effective so that candidates should be encouraged to make them. A review of the implications of changing answers by Mueller and Wasser (1977) concludes that "In general, students who change their initial answers to objective test items after deliberation gain more points than they lose by so doing. For every point lost by changing an answer, two or three points are gained, on the average." The authors further state that "From these findings it seems clear that students taking objective tests should be encouraged to deliberate on items for which they are not sure of the correct answer, and to change answers if, after deliberation, their first choice appears wrong."

Assuming that revising can increase scores, a second question is whether candidates actually convert the time for review into efforts to increase scores. This, of course, is another example of the intersection of the topics of speed and testwiseness. Little is known of the behavior of candidates during examinations. In the course of the present study, the project director had opportunities to observe three test administrations in three programs: the College Board SAT, the Secondary School Admissions Test, and the Law School Admissions Test. In each case it was apparent that some candidates finished early but did not undertake any review. Instead, they simply sat and waited for the time to elapse. For some, as the estimates above suggest, the time of waiting was 40% or more of the available time.

A third consideration is suggested by the discussion of speed in the ETS Test Analysis Manual, and by occasional reports from proctors and center supervisors. When large numbers of candidates finish early, they may begin to become restive, creating distractions for those still working and difficulties for those responsible for orderly test conditions.

The magnitude of this problem is seldom directly examined. At most centers, there simply are no overt problems of sufficient magnitude to lead to a report. But clearly, if work rates are normally distributed, the ETS standard that a test be completed by 80% of the group implies that many persons will have substantial blocks of time remaining after they finish the test. It would perhaps be appropriate to have the time-demand properties of the individual items estimated, and specific effort be made to introduce time-demanding activity at the end of the test, but it is doubtful that items could be inserted that were congruent with the earlier items and yet sufficiently time-demanding to "absorb" the time remaining. A later discussion in this report centers on the time requirements for items. It suggests that the range of differences among items in terms of the time they require is presently so small that the introduction of significantly more time-demanding material would be a challenging task.

The problems of optimum candidate behavior and optimum administrative efficiency obviously intersect. If candidates devote time remaining to a rereading of items, it reduces the administrative problems but increases the likelihood that time remaining will influence score differences. More adequate empirical information is needed on the factors governing candidates' decisions to review and on the efficiency of the reviews when they are undertaken.

Reports from Center Supervisors

As a regular part of each ETS test administration, the supervisors of the test centers are invited to submit Supervisor's Comment sheets. On these sheets, the supervisors offer comment of any kind, positive or negative. The Center Relations department at ETS summarizes these comments, offering a brief overview memorandum. As a part of this study, the Summary Comment sheets from most major ETS test administrations in academic year 1976-77 were reviewed. These comments might be anticipated to bear on such time problems as too many candidates finishing work. In general, the comments centering on time problems were simply too few for meaningful analysis. Occasional comments would be specific, as in the suggestion that "Section II of the test has too short a time limit." But in general, the major run of problems relate to the receipt of materials and the processing of candidates, and not to the actual administration of the tests.

The GRE program was no exception to this general pattern. The five administrations reported in the summaries led to 217 comments. Of these, thirteen dealt with timing. One suggested a shorter time for Section II of the Aptitude Test (October 1976) and 12 suggested that supervisors of the Advanced Tests be allowed to release candidates at their discretion during the last twenty minutes of the test. Of these 12, 11 were received after the December, 1976 examination, and one after the January, 1977 administration.

The evaluation of the implications of such forms a complex task. Why the number of suggestions rose following the December, 1976 examination is unclear. It is difficult to attribute the increase to specific difficulties with large numbers of restive, no-longer-working candidates, although the possibility cannot be ruled out. Perhaps a specific survey of proctors would shed light on the problems they face.

Time requirements for items

The test of linearity for the Reading subtest of the GRE Verbal XGR revealed an interesting pattern. The plot is presented Figure 9. In all, 14 items showed dropout on this test.

The first seven items to show dropout establish a roughly linear pattern. The last seven also exhibit approximate linearity, not quite parallel to the first. But the discontinuity reflects an interpolated reading comprehension passage. The test consists of passages followed by items. The graph clearly reflects the fact that the items per minute work rates may be highly similar for reading items in general, but the need to read the passage creates a quantum shift in the curve, a displacement to the right.

This finding is of interest because it shows that candidates in general probably proceed on the test by reading the passage in its entirety and then doing the items. If they proceeded differently, say by reading an item and finding it in the passage, the discontinuity would probably be less and would be followed by a somewhat nonlinear pattern, as familiarity with the passage increased through the processes of answering questions. Some candidates undoubtedly do proceed in this different manner but the general trend of Figure 9 is to the contrary; candidates tend to read the passage first.

Additional evidence of such discontinuities in reading material is evident in Figures 10 and 11, which present data from SAT reading comprehension pretests. In each case, the characteristics are similar: roughly linear plots within item sets following a passage, roughly parallel slopes to these lines, and displacements where the passages occur between items.

The plots then, can be sensitive to the time demands of material. There are approximate regularities in terms of increases in percentage failing to complete, from item to item. This regularity is disturbed by any unusual time demand, and the percentage dropping out is increased beyond the expectation. Items requiring less time than others would tend to show a vertical displacement in such a plot.

The interpretation of plots in terms of real time intervals derives from the use of these plots to estimate the mean and standard deviation of the total distribution. Assuming a basic normality, the parameters of the distribution

N
I U
T M
E B
M E
R

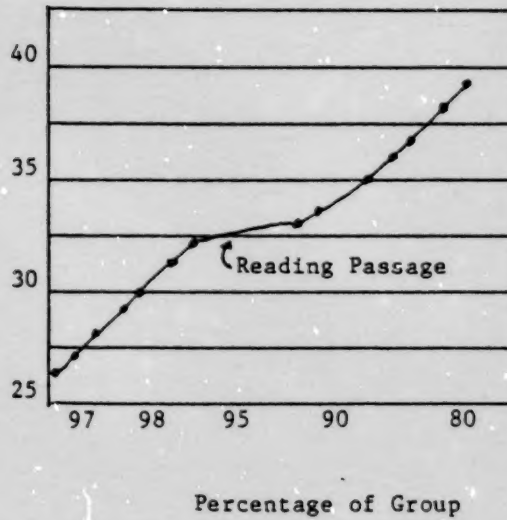


Figure 9. Percentage still working XGR Aptitude Reading Subtest.

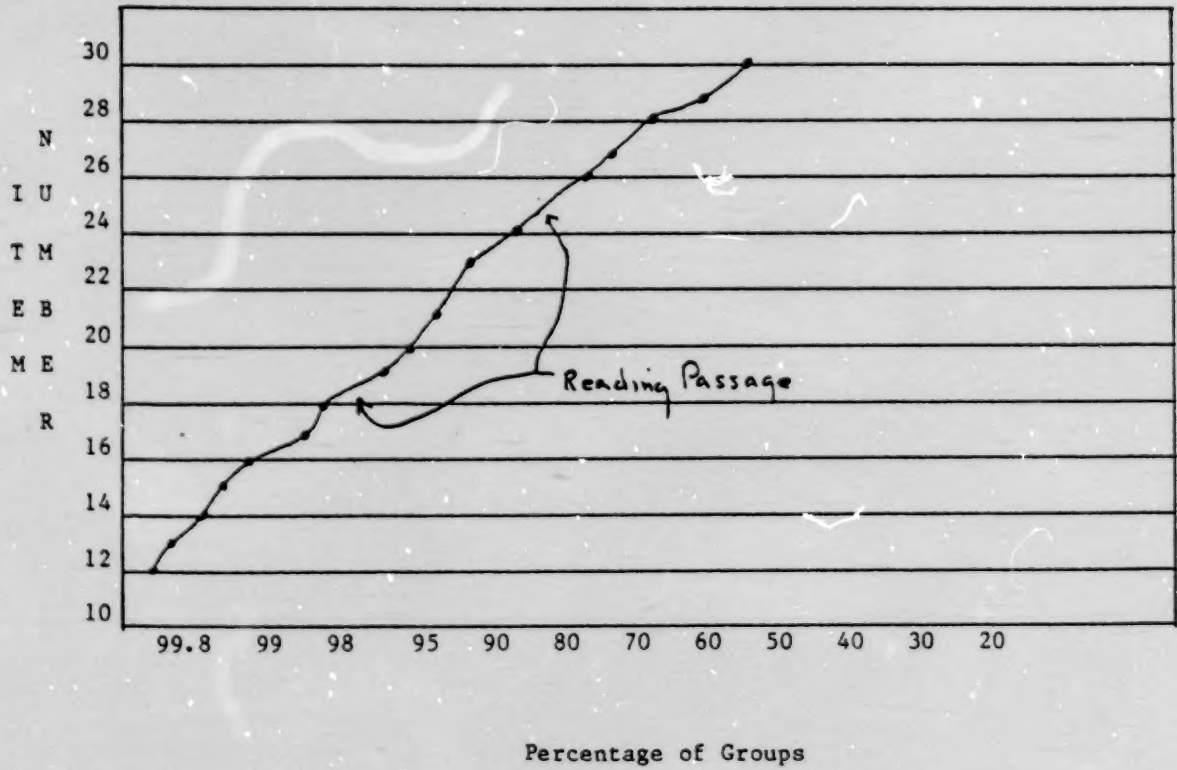


Figure 10 Percentage of groups still working - SAT pretest - Reading Comprehension RSA45₁₅.

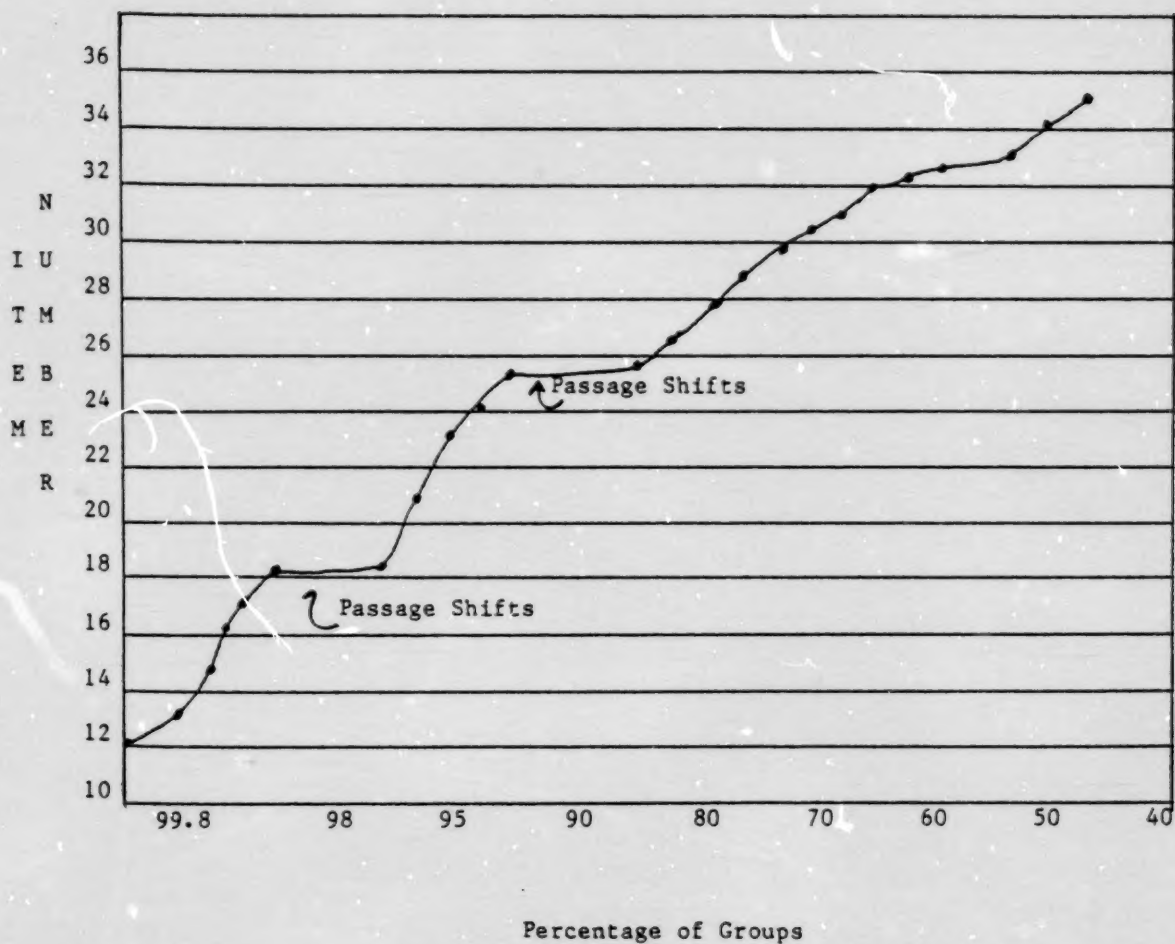


Figure 11 Percentage of groups still working - SAT pretest - Reading Comprehension RSA45₁₆.

can be estimated by drawing a line of best fit through the points and extending it through the 50% vertical. The number of items associated with this vertical is the mean of the overall distribution, whereas the standard deviation of items attempted is the difference between the 50% point and the 84% point.

The general logic of these methods can be used to infer the time demands for the individual passage. Figure 12 shows the plots for Figure 11 redrawn. Parallel lines are imposed on each of the first three item sets. These lines obviously do not fit any given item set very well, but they are a general synthesis of all the items. The lines are used to infer vertical distances, A, B, and C. These distances are interpretable in terms of numbers of items, using the vertical axis of the graph. Thus, line A indicates a vertical displacement of 7 items. Given the experience of the set of points which precedes A, one would expect the dropout associated with item 19 to be associated with item 26. It is as if the candidate, in reading the passage, took the time equivalent to that spent in answering seven items.

This time can be estimated. The average candidate, in terms of the line which precedes A, can complete about 51 items in the time allowed, 30 minutes. This is about 36 seconds/item. Applying this estimate, the passage determined by A requires $7 \times 36 = 252$ seconds or a little over four minutes. B is the equivalent of approximately 4 items, or about 2-1/2 minutes. C, equivalent to 3 items, is perhaps 2 minutes.

These passage time estimates account in all for about 8-1/2 minutes. The graphs suggest that candidates actually lost 10 minutes on the average since their average number of items completed is about 34 versus the 51 it is estimated they could complete.

The data are presented to demonstrate the method, rather than as precise descriptions of the time requirements. In particular, it should be noted that the passages which precede the items which define the estimating line have actually consumed time, and have reduced the estimate of the rate at which candidates process items. The data suggest, however, that the time requirements of material can be estimated through the dislocations of item plots.

Figure 13 demonstrates the general application of the method to the sentence completion data described in Figure 2. The effort to plot a family of parallel lines and assess the vertical distances is in general unsatisfactory. The distances are too small to be interpreted, given all the sources of error in the plotting and visual linefitting. The average sentence completion item may be estimated to require 33.3 seconds for the average candidate. There is a suggestion here of possible displacements as great as this, so that an item may be requiring upwards of 60 seconds, but the general pattern is one of considerable equivalence among items in time demand.

The application of this general method for determining the time demands of material to the lesser irregularities associated with individual items is

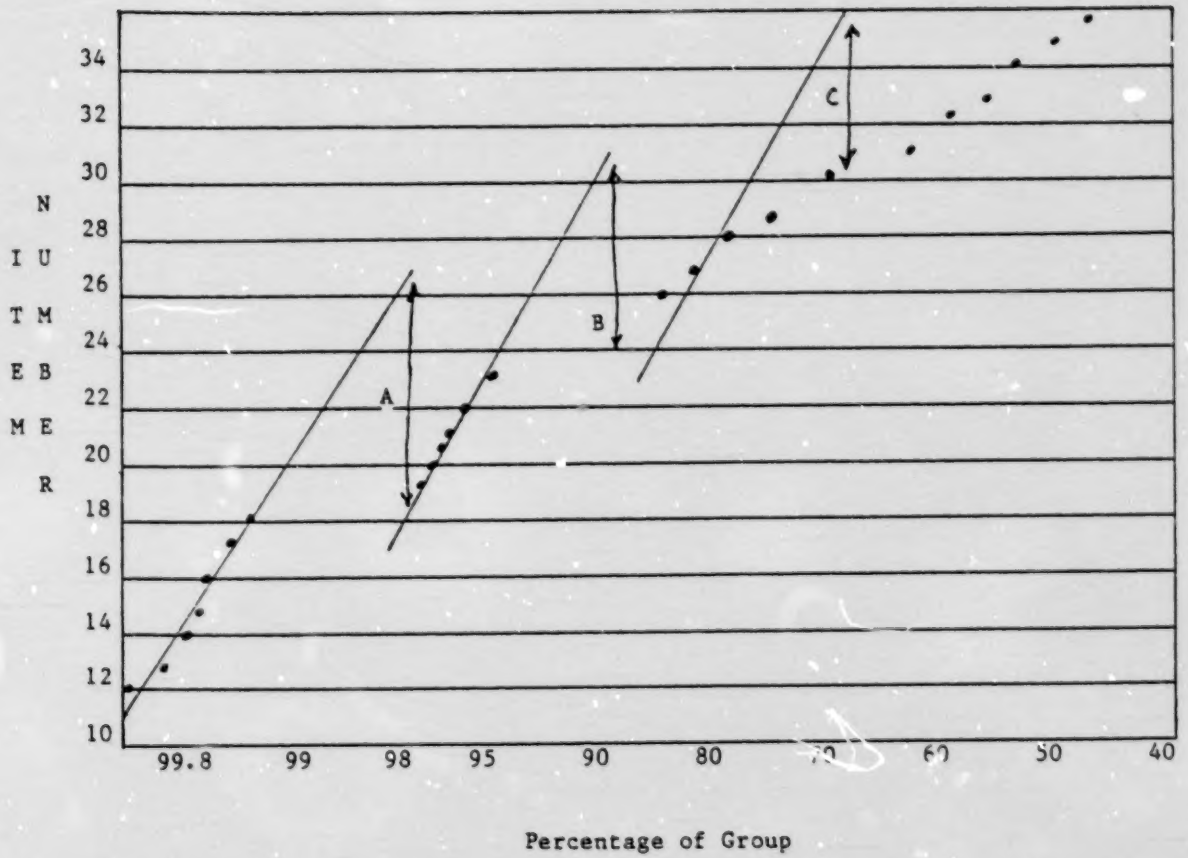


Figure 12. Estimated time required through SAT pretest - Reading Comprehension RSA45₁₆.

quite uncertain. The plot in Figure 2, for example, shows some minor irregularities. But the estimates of time required would be in terms of seconds, a demanding estimation procedure for a method already burdened with approximation and assumption. More empirical studies would be needed before the reasoning which estimates the passage time requirement is successfully applied to the level of the individual item.

Within broad limits, however, the plots reveal information about the average time demands of items. The pattern they reveal is one of little variation. There is no evidence that any individual item requires anything remotely like the time required by a passage. Here and there, there seems to be evidence of items that have more or less than the anticipated dropout, but the variation is very small.

One of the major implications of these data, then, would seem to be that test items as they are currently produced have a fairly constant average time demand for the candidates. While the interpolation of a time-consuming activity such as a passage is detectable as a kind of sideways translation of the points in the plot, there is little evidence of the effects of variations in the time demands for individual items.

This evidence of considerable similarity in the time demands for items of a given type is a little remarkable in that there is almost no formal attention to time demand in the evaluation of individual items in the test development process. The test specifications do not refer to time demand for individual items in any way; the pretest tryouts and the judgments of test specialists are silent on this point. Virtually the only attention formally given to time for specific material within a test comes from the concern when an overly-long reading passage is suggested.

A first conclusion from this lack of variation might be that there is little evidence of a kind of potential test bias in which some candidates lose score points because they tackle an extremely time-demanding item while the other candidates omit it and move on to other items which require less time. The lack of variation in item time demand is clearly a guard against this kind of bias.

But the relative constancy of time demand may also reflect the mental processes these items draw upon. That is, the evidence that the production process in fact produces items with remarkably consistent time demands may be a reflection of some inherent similarity in the solution processes demanded by the items. Although the various item types differ among themselves in average time demand, there is little difference within: one verbal item or one quantitative item is very much like another. What is involved in them is a reading, recalling, and reasoning process that probably does not vary greatly from one item to another. If the components of the process are present in the respondent, in the sense of the required reading, recalling, and reasoning abilities, the process takes very little time. Even erroneous process has a limited variation.

Opponents of multiple-choice testing often deplore the kind of process simplification which these data imply. Real world problems are more varied, complex and demanding, it is argued; there is a trivialization in the reduction

Item
Number

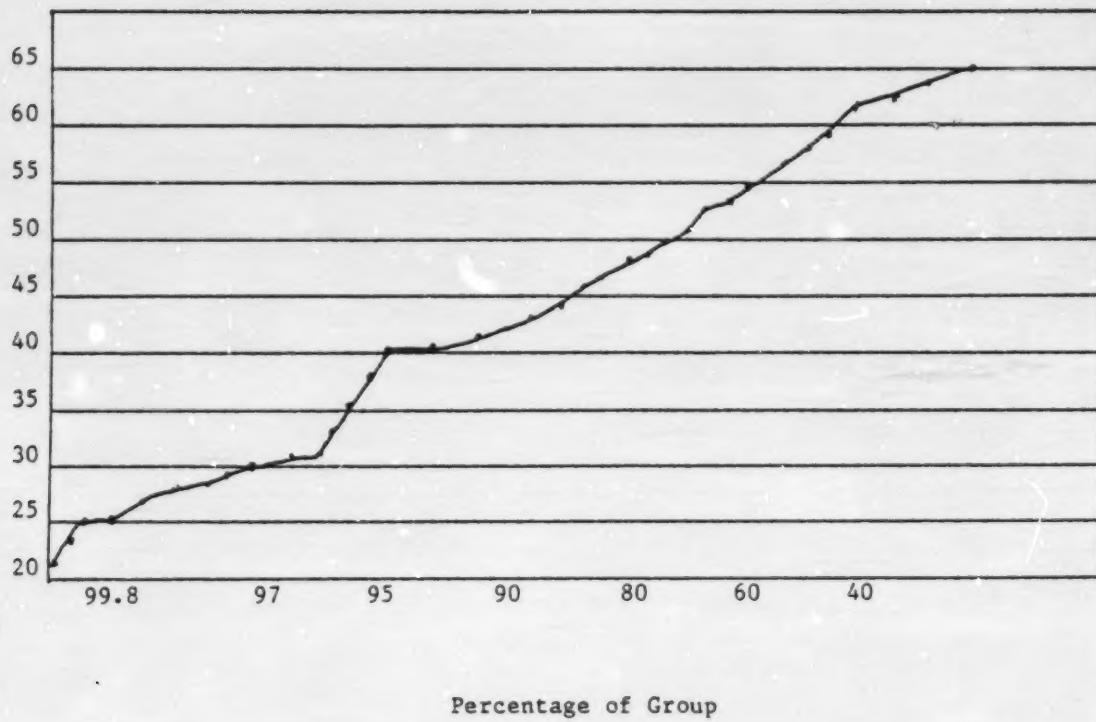


Figure 13. Percentage of group still working - SAT Sentence Completion RSA45₁₇.

of the process to the rapid selection among alternatives. A notable example of such criticism is that of Hoffman (1962) in his book "The Tyranny of Testing." It is true that the solution processes to multiple-choice questions have seldom attracted research attention. Brigham (1932) in "A Study of Error" was a major exception. Connolly and Wantman (1964), Bloom and Bloom (1950), and Gentile (1966) were others. The general lack of attention, however, may reflect the very narrowness of process that is reflected in the time data. Candidates cannot report the process effectively because it is so limited an exercise in thinking that it lacks dimensions or structure.

The move to an evaluation of process implications is admittedly premature in a context where the meanings of the item plots themselves are not unambiguous. But the general evidence of the average time required by a test item is that it is typically under a minute, and for basic verbal knowledge and reasoning items, only 20-30 seconds. The processes demanded in this short interval cannot be very complex.

A Limited Empirical Check

The concern for these process implications led to some limited testing of the time demands of individual items. The present study could not collect adequate evidence on the time demands of individual items. However, a small group of graduate students was asked to react to GRE Aptitude items under timed conditions. Each item was reproduced on a card. The card was presented to the subject, and a stopwatch used to time the interval until response. Approximately 10 students were used for each item, although this number and the composition of the sample, varied from item to item.

The items used were Verbal items: Sentence Completion, Antonym, and Analogies. They were classed as Easy, Medium, and Hard, according to the ETS difficulty estimate, delta. Items were selected as Easy if they showed a delta of 8.0-9.9; as Medium if they showed a delta of 12.0-13.9, and as Hard if they showed a delta of 16.0 or greater. There were four items in the Easy and Middle categories, for each item type. There were no Hard items in the Sentence Completion category. The number of Hard items was brought to 12 by selecting six Antonyms and six Analogies for this difficulty level. The results are presented in Table 6.

Table 6: Average Time/Item for Three Verbal Item Types

Item Type	Difficulty Level	Seconds/Item
Sentence Completion	E	16.0
	M	25.3
	H	----
Antonym	E	5.0
	M	10.8
	H	16.7
Analogies	E	16.0
	M	34.3
	H	35.3

These data suggest that there are differences in time demand, both between items and between difficulty level within items. Thus the results do not confirm the homogeneity of item demand reflected in the plots. But they cannot rule it out, either. The small numbers of very able students who participated are in no sense representative of the GRE population. The very small sets of three or four items are obviously a limited sampling of GRE Verbal content. Nonetheless, there is the suggestion of regularity in relationships, and these would appear to be consistent with the intuitive expectations derived from assumed solution processes. That is, in terms of solution processes, the item types would be ranked from simple to complex, as Antonyms, Sentence Completion, and Analogies. The time demand data seem to support this.

These data also bear on the interpretation of the linearity of relationship observed in the normalizing plots. Since the more difficult items are typically placed at the end of a test, one might expect a slowing down of the group on these items. This would be consistent with Table 3. This would produce a bunching of items attempted scores, and a consequent distortion of the plot. In general, this does not happen either because dropout affects only the last few items and these are more homogeneous, or because speed and power are related, and the persons who continue to work are faster and more able. A full determination cannot be made without a more careful study. It is a paradox, however, that one can so easily demonstrate item time demand differences by direct administration to students but not confirm these in the graphic item plots.

The relevance for test development of findings such as are suggested here is not direct. Establishing empirical evidence of different time demands does not contribute to any revision of the test specifications. It is evident that greater test reliability could be secured through more items, and that more items can be introduced if the time demand per item is reduced, but it cannot be seriously suggested that the content of tests be constrained so severely in order to achieve a greater time efficiency. The methods may have promise in safeguarding against overlong passages or in studies of test wiseness.

Omissions and Speed.

The ETS analysis of speed focuses on the number of Not Reached responses as the index of subject rate-of-work. As explained earlier, this analysis ignores omitting, or considers an omission a reasoned decision about the item, requiring a fairly complete reading of the item and approximately as much time as any other response. By this reasoning, omissions are not considered an index of speed.

The Gulliksen analysis essentially never confronted the omit response and its implications for speed. The analysis lumped as "unattempts" any failures to provide a response, whether through omitting or failing to reach. Gulliksen recognized the limitations of his ratios for tests which depart significantly

depart significantly from pure speed or pure power, and under either of these "pure" conditions there are no omissions. To the extent that Omits occur, of course, they contribute to an appraisal of the test as speeded, in a Gulliksen analysis.

Cronbach and Warrington's approach does not directly consider the Omit. Instead, it is subsumed into the larger question of the underlying comparability of the true scores for Total Right under speed and power. Essentially, this reflects the fundamental Cronbach and Warrington emphasis on the evaluation of speed as consistent or inconsistent with power, rather than as a quantity of some type to be described in an absolute sense.

Stafford's approach is basically like Gulliksen's, in that it computes the measurement of speed by considering the class of "unattempted" items. But Stafford does not discuss the possibility of a distinction between Omit and Not Reached, nor evaluate the merits of such a distinction. As is the case with Gulliksen, Stafford essentially assumes that an "unattempted" item is a Not Reached item, or that at most the number of omitted items is not large.

The Reilly-Donlon indices derive from the ETS analyses, and rely upon a Not Reached score as the fundamental measure of speed. In all, then, the question of speed and Omitting is not directly confronted by any method. The two topics, however, are logically intersected if one hypothesizes a model for the Omit decision other than the "full consideration" model. Under the "full consideration" approach an Omit takes as much time as a marked response. But if this is not true, if some individuals can process the Omit decision more rapidly, they will proceed farther through the test than they might have. If speed is the extent to which test scores reflect distinctions among individuals in rate of work, then the Omits score will reflect speed under these conditions. People with equal Not Reached scores but unequal Omit scores will not have performed at the same rate of work.

Comparing tests with regard to the levels of Omitting they generate is rendered difficult by such factors as differences in test length. The data in Table 7 present some representative GRE program data. The basic ETS speededness data are presented, also.

The mean levels of Not Reached do not vary greatly across tests. The exceptions to this tend to be Form WGR Economics, which showed an unusually high level of not reached and the Engineering forms which revealed a somewhat similar pattern. Biology, Chemistry, Psychology, Education and Computer Science all show quite small levels for the NR average. Expressed as a percentage of the length of the test, the range of values is from virtually zero to about 5% for a few forms.

The level of Omitting tends to be more substantial than the level of Not Reaching. The proportions of the total number of items which are omitted, on the average, range from .11 to .24, with most tests falling in the 15-20% range.

TABLE 7

Mean Number of Omitted Items for Selected GRE Tests,
with Related Speededness Data

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Form	Test	Total Items	Total Time (min)	M Omit	M NR	% Comp Test	% Comp. 75% of Test
WGR	BIO	210	(170)	29.41	0.79	92.7	99.7
XGR	BIO	210	(170)	27.90	1.64	74.6	99.9
YGR	BIO	210	(170)	32.93	0.70	90.3	99.8
VGR1	CHEM SEC I	55	(60)	10.28	1.81	58.8	98.1
	II	55	(60)	12.76	1.37	72.0	98.2
	III	20	(25)	2.70	1.07	49.0	96.8
	IV	20	(25)	4.58	0.76	62.1	98.6
WGR	CHEM	150	(170)	33.38	2.17	75.6	99.1
YGR1	COMP SCI	80	(170)	16.34	1.29	62.1	99.7
WGR	ECON	170	(170)	29.05	8.14	26.3	96.5
VGR1	ECON I	60	(60)	6.43	1.73	67.4	98.4
	II	60	(60)	9.87	2.21	71.4	95.2
	III	25	(25)	3.16	0.95	76.4	96.9
	IV	25	(25)	3.95	1.17	74.7	96.1
XGR	EDUC	200	(170)	29.16	1.36	86.0	99.7
WGR	ENGR	150	(170)	35.55	5.85	57.4	98.1
XGR	ENGR	149	(170)	36.67	7.20	20.4	98.1
XGR	PSYCH	199	(170)	31.75	0.69	65.8	99.9
YGR	PSYCH	200	(170)	28.42	1.17	58.0	99.5

The ETS analysis for speed introduces Omitting into its considerations somewhat obliquely. The cross-tabulations for score versus attempts ($R + W$) necessarily reflects Omitting ($O + NR$) since $(R + W) + (O + NR)$ equals N by definition. In the ETS analysis, each column is a different level of attempting. To the extent that there are individuals at different levels of attempting but achieving virtually 100% success, on the items attempted, the analysis considers the test as possibly speeded. Essentially, this reflects the Gulliksen assumption that Omits are few, and that differences in total attempts ($R + W$) reflect differences in NR level. As the data in Table 7 show, the Omit response is typically a far greater determiner of total attempts than the Not Reached response. In Table 7 the mean number of omits, in column 5, is considerably greater than the mean number of Not Reached responses in column 6.

One principle underlying this aspect of the ETS analysis is essentially as follows: if two persons each exhibit 100% success on items attempted, but they attempt different numbers of items, then there is a potential for a marked speed component in their score difference, for the person with fewer attempts might continue to be highly successful, if given more time, and erase the score difference. If the group under analysis exhibits sizable numbers of persons with virtually perfect scores but widely varying number of attempts, the test analyst will judge the test speeded.

This ETS principle is susceptible of some generalization. Whenever two individuals enjoy similar success rates on the items attempted, but a marked difference in the number of attempts, the effect of more time could be a reduction in the observed score difference. That is, the ETS analysis focuses on individuals who succeeded on virtually every item attempted, but the problem exists for other levels of success as well. Only if rate of success is somewhat correlated with rate of attempts is there evidence that the standing of individuals might not be changed with more time.

Some confirmation for this generalization is provided in the ETS Test Analysis Manual where Figure 14 is presented and accompanied by the following discussion:

Some candidates confronted by an unusually difficult test apparently are willing (or forced) to 'give up.' And when a candidate gives up he does one of two things: either he stops working and thereby gets a high 'not reached' score, which can be misinterpreted by the analyst as an indication of speededness, or else he marks answers to all the rest of the items without any serious attempt to answer them correctly--the last portion of his answer sheet may even show a system of 'pattern marking.' If the test is so difficult for the group as a whole that a substantial portion of the candidates follow the second procedure, then the mean score corresponding to the right-hand column of the two-way distribution may be a little lower than means corresponding

to columns immediately to the left of it. Form VGR of Advanced Spanish (SR-73-97) provides a good example of this kind of candidate behavior. Another example, taken from the language study reported in SR-57-16, involves data for a French Achievement Test analyzed separately for 2-year, 3-year, and 4-year French students. Mean scores corresponding to successive columns are plotted in Figure [14]. The 4-year and 3-year groups obtained higher scores, on the average, the more items they attempted. For the 2-year group, however, the test is very difficult: the average score for those who responded to 110 to 125 items is lower than the average for those who tried 100-109 items. Since formula scoring is designed to eliminate from the earned score any effects of random guessing (on the average), this finding suggests that low-ability examinees tend to answer difficult items that high-ability examinees are more likely to omit. The detailed item analyses frequently exhibit support for this theory.

Inspection of the graphs in Figure 14 obviously confirms the suggestion that the two-year students failed to show the score increases which three-year and four-year students demonstrated for increased numbers of attempts. But there are other characteristics of these curves that may be potentially relevant to an analysis in terms of speed. Both the three-year and four-year curves show a tendency toward "acceleration," toward greater score-per-attempt ratios as more items are attempted. That is, those 4-year students attempting 80 items earn approximately 43 score points, on the average, or about .54 score units per attempt. Those 4-year students attempting 120 items have average scores of approximately 83 or about .70 score units per attempt. For the three-year students, the score per attempt is about .42 for students with 80 attempts, but .65 for students with 120 attempts. The two-year group exhibits some similar increases prior to the peak value at approximately 105 attempts. Thus, for 60 attempts, two-year students show score/attempt ratios of about .30, for 100 attempts, about .40.

These score/attempt ratios may be thought of as the efficiency with which a subject achieves score points. A low score/attempt ratio is inefficient, a high ratio yields greater rewards. What pattern of efficiency might be expected for a speeded test, or a power test?

As the ETS test analysis manual suggests, evidence of sizable numbers of persons with a constant efficiency of 1.0 but variable numbers of attempts is cause for concern about the effects of speed. In general, an unspeeded test should probably demonstrate increasing score/attempt ratios across the range of attempts. The failure to show such increases does not establish or confirm speededness, anymore than the ETS indicator does. One cannot unequivocally attribute differences in attempts to differences in speed when omitting is the major source of failure to attempt and where the relation of omitting to rate of work is not clearly understood. But constancy of the score/attempt ratio is a signal that those candidates with fewer attempts may be disadvantaging themselves. When the score/attempt ratio is .50, 10 items omitted may produce a five-point raw score deficit.

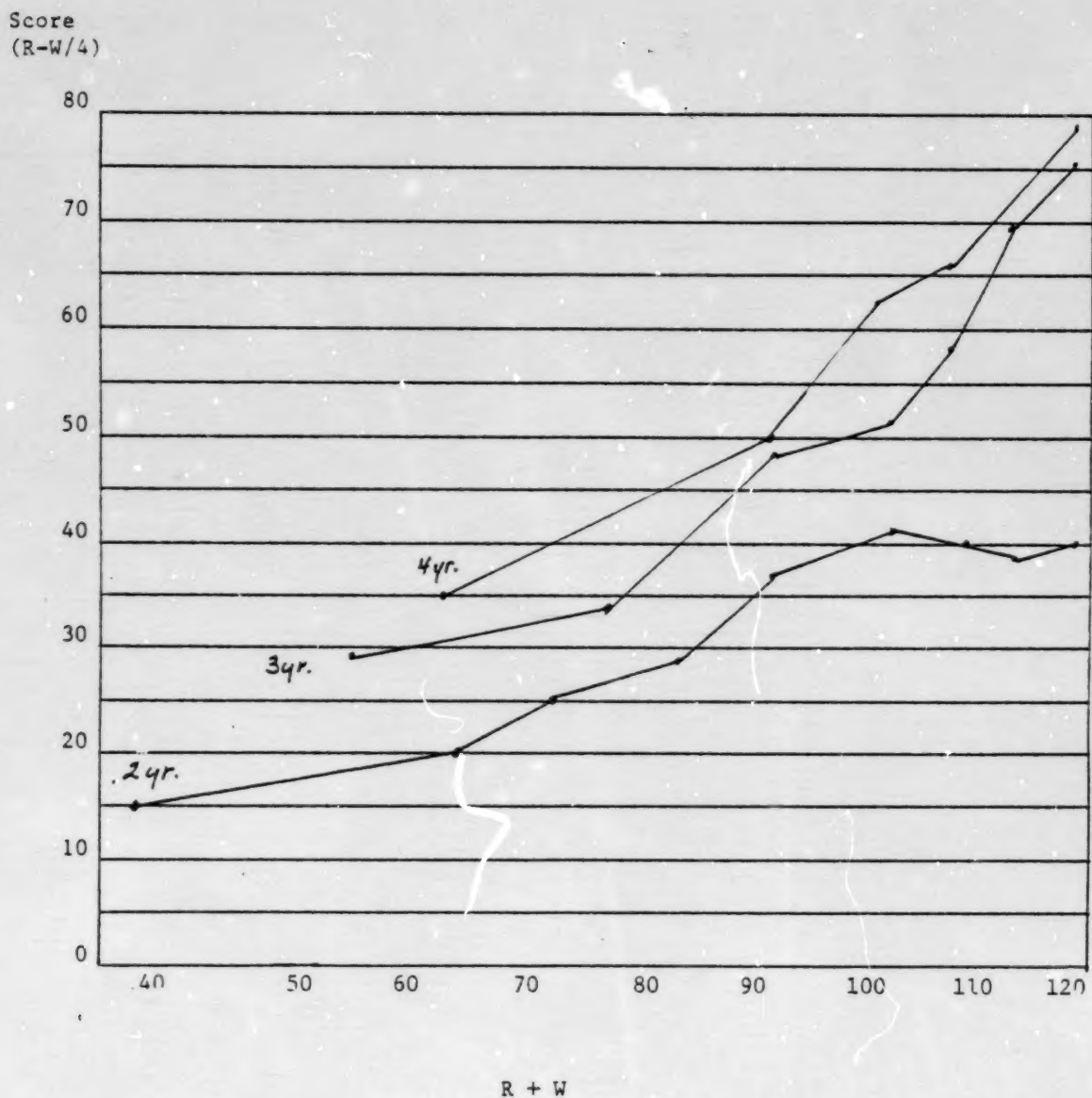


Figure 14. Mean French Achievement Test Scores corresponding to successive Levels of R & W for candidates with two, three, and four years of study (from Swineford 1974).

Figures 15-20 show plots of efficiency for various tests. Figure 15 is the plots of Figure 14 redrawn in terms of the "efficiency" scale. Figure 16 is a plot for GRE Engineering, with a suggestion of the increasing efficiency that reflects an appropriate speed-power correlation. Figure 17 shows considerable "noise" in the plots, but some suggestion of increase. Figures 18 and 19 seem quite complex, with somewhat abrupt declines in efficiency for those answering most items. Figure 20 tends to show consistently declining efficiency, though this cannot be established with certainty.

The interpretation of these curves in terms of speed is difficult. In general, they do not show rising success ratios with a larger number of attempts, so that the potential for score differences based upon rate-of-work differences is clear. The problem in interpretation, here and in the ETS indicator, is whether the score/ attempt ratio would persist if the subjects made more attempts. If it would, score differences would be diminished and the test as it exists has a speed component.

The problem of omitting is generally handled by assumption, in current speed evaluations. The principal assumptions typically are that the candidate's decision to omit is a reflective one, and, to a large extent, an appropriate one; that groups with different numbers of attempts show roughly equivalent success ratios because the less able group (in the sense of lower score) limit themselves to a selected subset of easier items. Thus, the apparently equivalent success ratios are, in fact, incommensurable, being based upon items of varying difficulty.

These assumptions could and should be tested empirically. The role of omitting in the assessment of speed here blends into the larger need to understand its role in test performance generally. That is, issues of fairness to candidates, and concerns for the effects of test wiseness argue for a consideration of the reasons for and outcomes of omitting. A review of some of these considerations, in the context of guessing strategies, is provided by Pike (1978), in concurrent and related GRE studies. The present level of understanding of Omitting and speed seems deficient. The major assumptions concerning candidate behavior: (1) equal time for Omit or response and (2) decreasing score/attempt ratios if candidates are given more opportunity or motivation to attempt are susceptible of empirical investigation.

The assessment of the second assumption, concerning the validity of candidates self-selection of items, needs empirical study, in particular, in view of the report by Mueller and Wasser (1977) that candidates achieve potentially valuable score/attempt ratios.

Implications for Program Practice

One major purpose of this study was to consider implications for program practice. In general, the review found no evidence of major and obvious problems with speed on any test. Such findings were not to be anticipated, since a formal

Efficiency

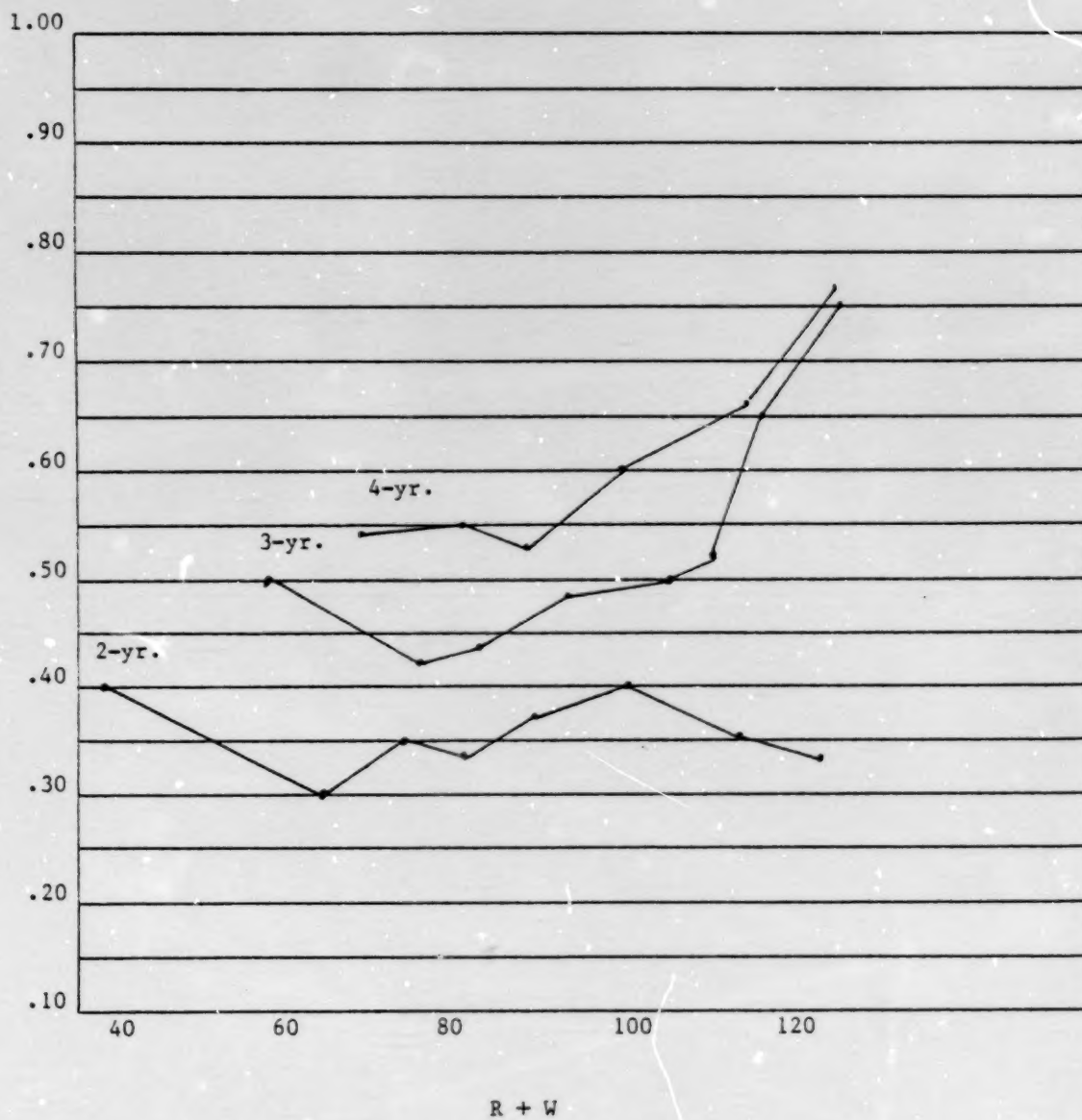


Figure 15. Efficiency of French Achievement Test performance at successive levels if R +W for candidates with two, three and four years of study.

Efficiency

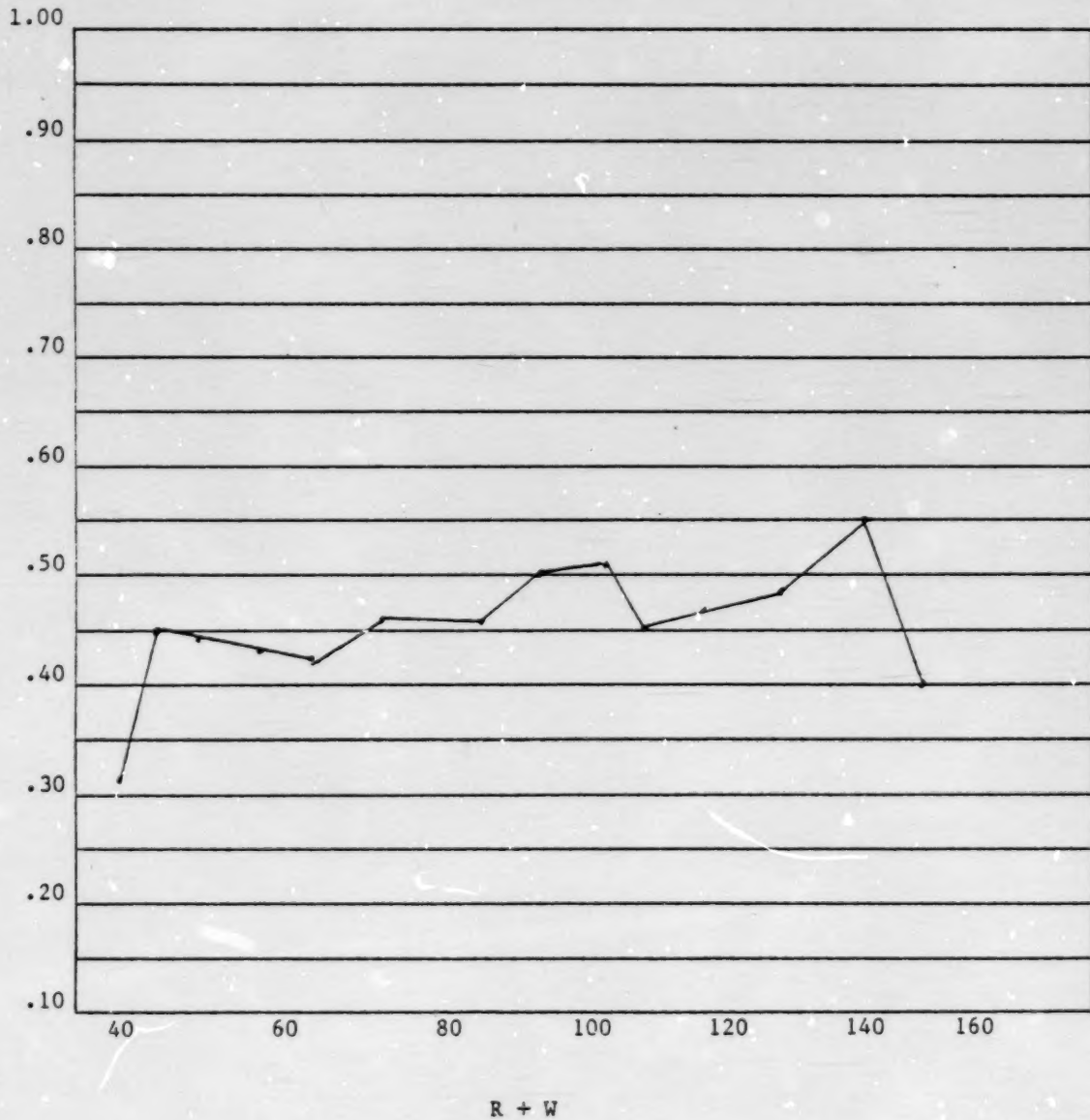


Figure 16. Efficiency of Engineering Achievement Test performance at successive levels of R + W.

Efficiency

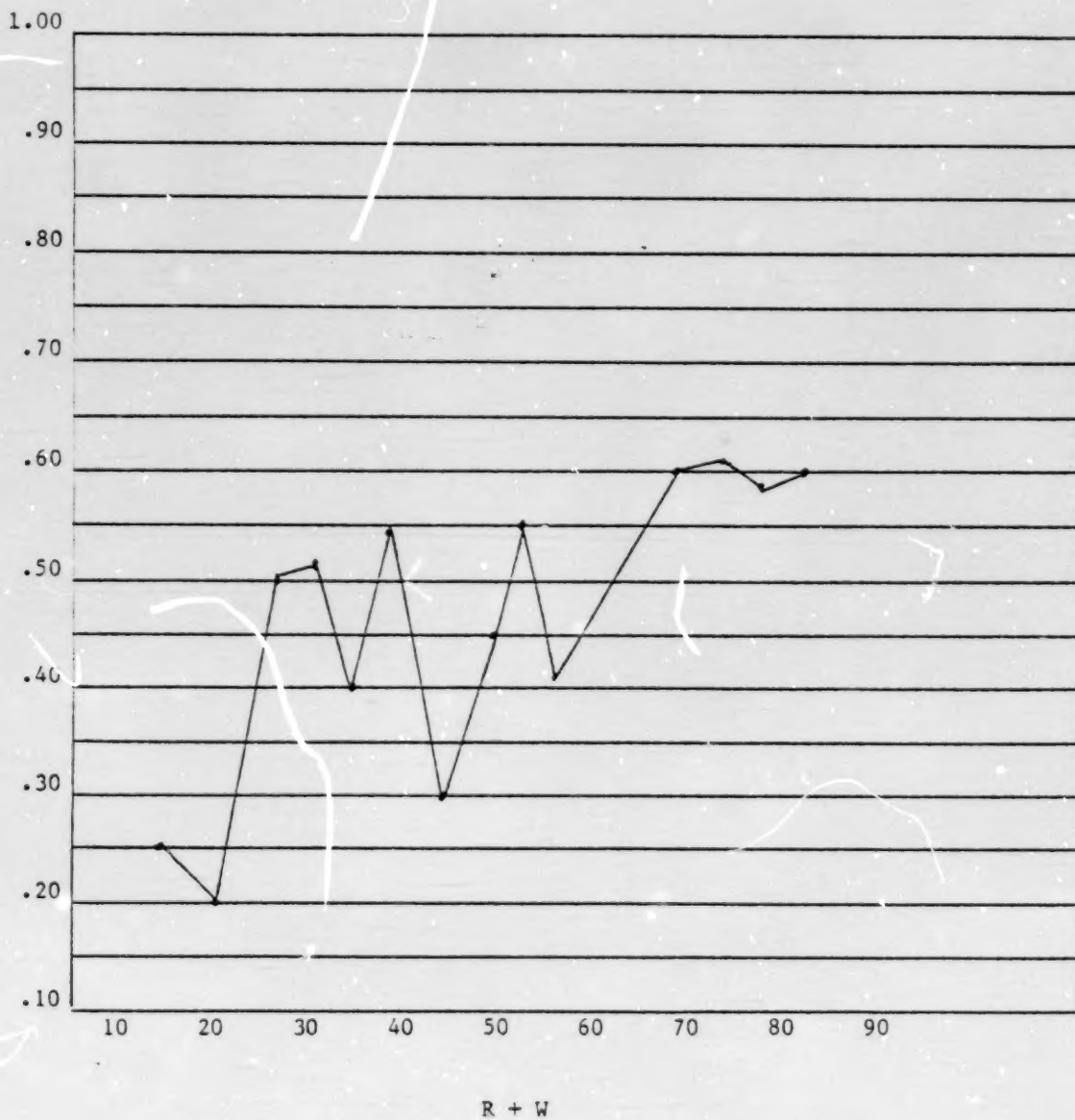


Figure 17. Efficiency of Biology Achievement Test performance at successive levels of R + W.

Efficiency

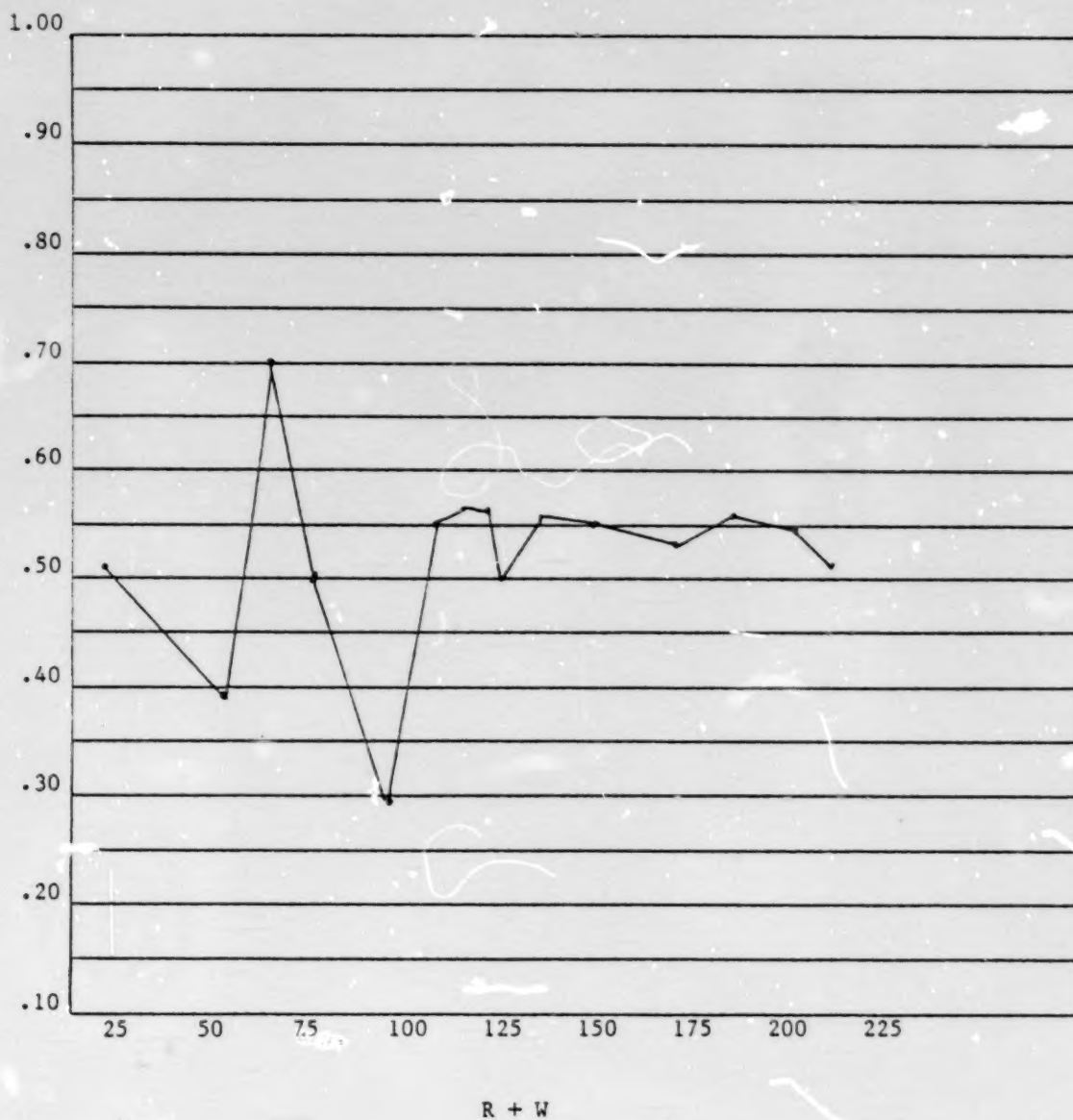


Figure 18. Efficiency of Biology Achievement Test performance at successive levels of R + W.

Efficiency

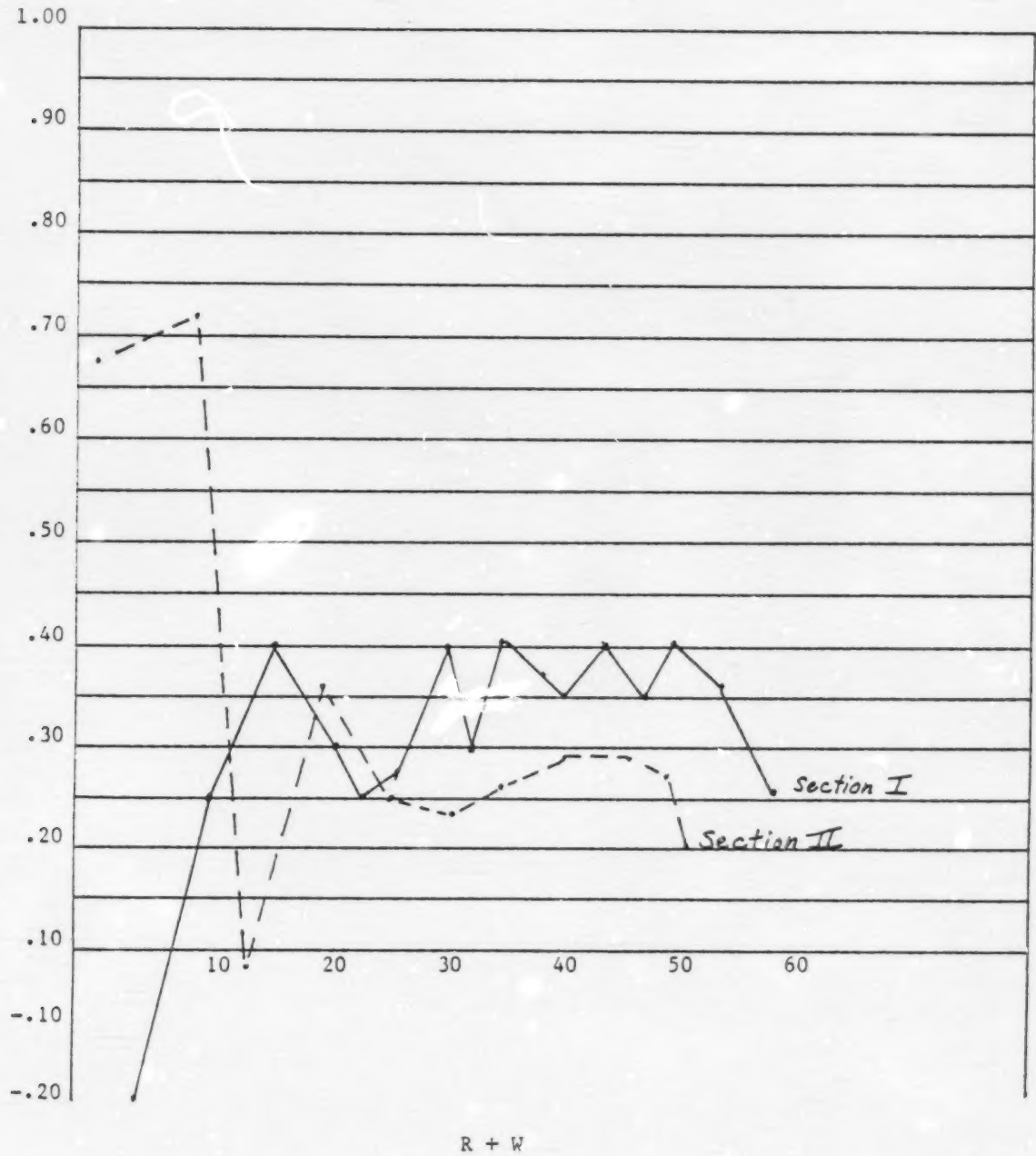


Figure 19. Efficiency of Chemistry Achievement Test performance (Section I, II) at successive levels of R + W.

Efficiency

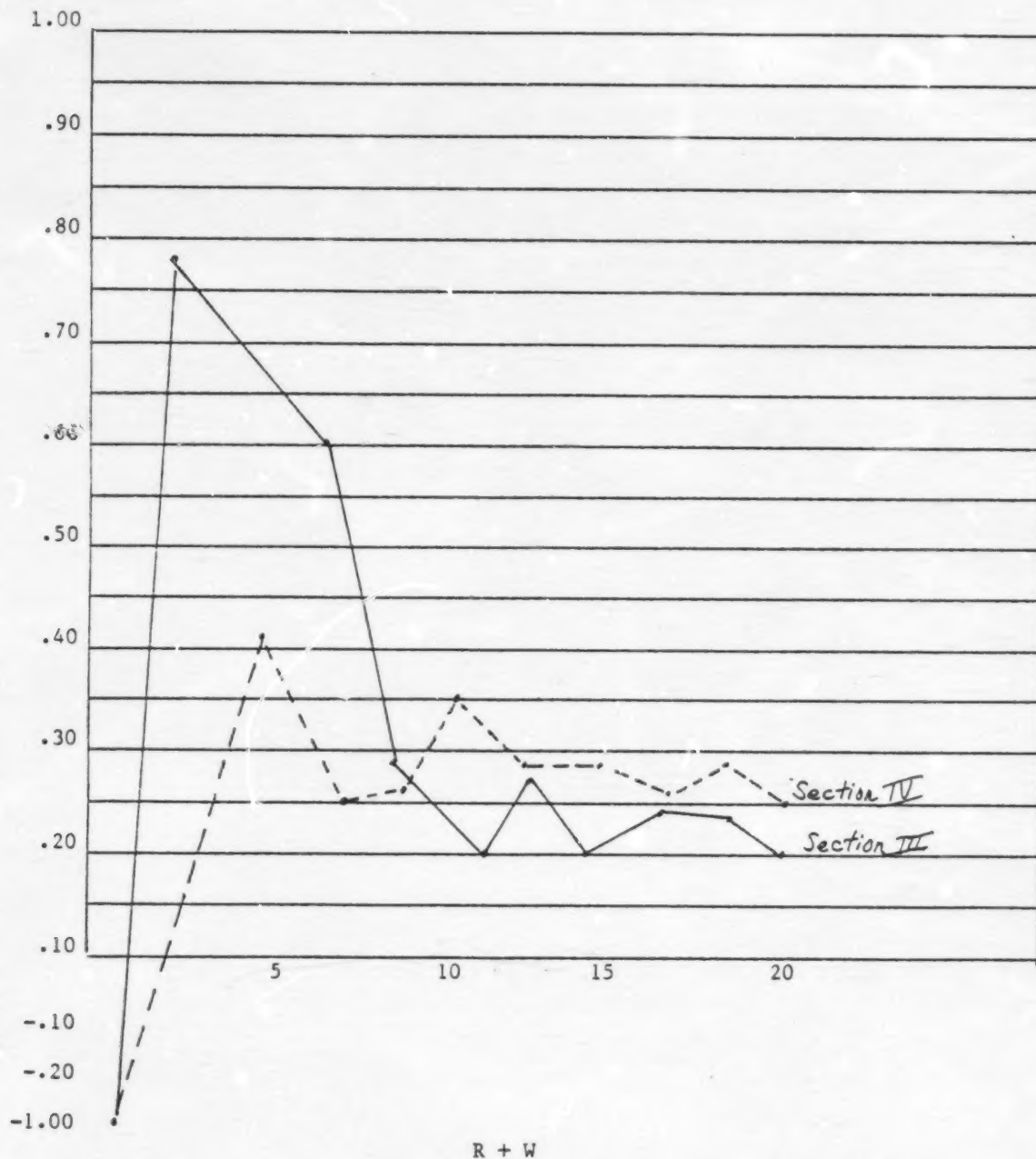


Figure 20. Efficiency of Chemistry Achievement Test performance (Section III, IV) at successive levels of R + W.

appraisal of speed has been a regular component of the test analysis, and excessively or inappropriately speeded tests would be detected by the ETS review procedures. However, the ideas and approaches suggested herein would seem to have practical potentials.

The following suggestions with regard to the description and assessment of speed are offered:

1. Consideration could be given to supplementing or replacing the traditional ETS test analysis presentation of data, such as Tables 1 and 2, with a graphic description of dropout such as that presented in Figure 21. The graphic presentation has the value of retaining everything currently provided by the three points included in Table 1 but adding a large amount of additional information in a swiftly reviewable manner. Not the least of its advantages is the ready availability of the dropout observed at the next to last items. It is intuitively clear that when the last items show much more dropout than their immediate predecessors, the use of this index must be abandoned. But charts like Table 2, by presenting a single point centered on the last item (100% of the test) place the burden of presenting the supplementary information in the text.
2. Standards for speededness could be developed in terms of the entire dropout plot. Essentially, this would consist of establishing a zone, within which dropout plots should remain in order to be considered speeded. Such a "Zone" is sketched for Figure 21.
3. If the assumption of normality for the work rates on the test are reasonably useful, the dropout plots could be prepared using normal probability paper in a manner analogous to those developed for this report. Such visual presentations would assist the users in judging the approximate tenability of the normality assumption in interpreting the kinds of one-administration speed-power correlations proposed in this paper. Both the Reilly-Donlon estimate and the biserial correlation estimate depend upon approximate normality for the reasonableness of their procedures.
4. The current test analysis offers no correlational data regarding the four basic scores: Rights, Wrongs, Omits, and Not Reached. This information is available based on the analysis samples of candidates, in the computer output for the test analyst. It is sufficiently useful in the appraisal of speed that it could be presented regularly. The correlation of Rights and Not Reached or of Omits and Not Reached are intrinsically relevant to the appraisal of speed.

Percentage
of group

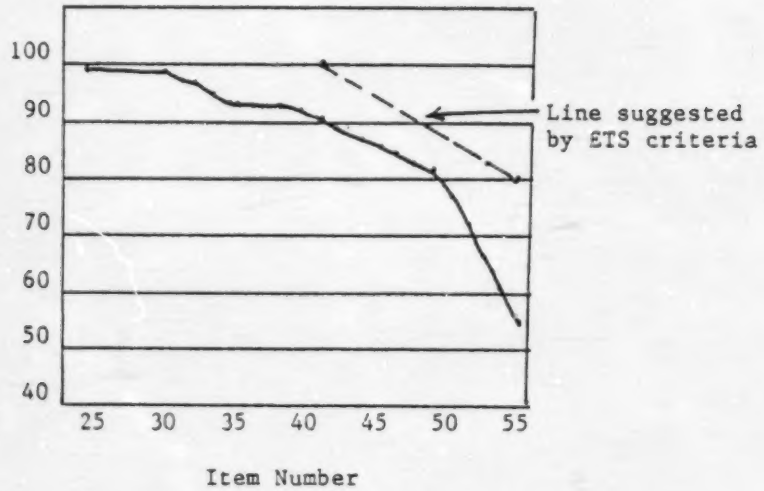


Figure 21. Sample graphic for describing speed.
VGR2 Vocabulary Subtest.

5. If the research into the usefulness of the one-administration correlation coefficients proves promising, the reporting of such coefficients could be made a regular element in the test analysis data relating to speed. The advantage of these specific estimates lies in their generality and convenience. The general utility of the speed-power correlation is that of being a useful indication that any speed variance is not unrelated to the power measure.
6. The current contingency table for (R + W) versus score (as in Table 2) could be usefully supplemented by a graph of the regression of score on this variable, in the manner presented in Figures 15 - 20. These graphs could be an important element in the understanding of the efficiency of candidates who work at different rates of speed.

Implications for Further Research

One purpose of this study was to identify areas of potentially useful future research. The bibliography presented with this report suggests the many possibly relevant lines of research which might be considered. The following list, then, is selective and intended to lead to the principal goals of a) providing measurement specialists with adequate information upon which to base an appraisal of speed, b) providing candidates with the best counsel as to how to take the test, and c) providing test developers with information on how to control or modulate the speed characteristics of tests.

1. A formal study of the assumption of normality of work rates would seem to be useful, primarily as a means of extending the empirical confirmation of the assumptions presented in this report.
2. An investigation of the extent to which one-administration coefficients are useful estimates of the empirical speed-power correlations they are analogous to would be helpful. Both the Reilly-Donlon methods and the biserial estimates are estimates of correlations which would be observed in data where everyone had time to complete the number of items in the test, and anyone could work on additional items if their work rate was such that they would normally have some time remaining.
3. An investigation of the extent to which rate of work is invariant under similar basic instructions but with different time limits and numbers of items. This research is relevant to the evaluation of the generality of the findings of test speededness, and to the kinds of suggestions which can be given to test developers as to formal strategies by which they might influence candidate behavior in ways that are considered desirable.

4. A study of the time demands of individual items. A minor survey of the potential for differential time demands by individual items was undertaken in the current study. The essential motivation for this derived from the impression of an apparent equivalence of time demand among items which emerged from the normal probability plots. That is, while a major interpolated activity such as a reading passage produced a noticeable effect, the general impression was one of limited effects by the individual items.

The fundamental data envisioned for the study suggested here are more precise estimates of the average time required by different types of items by different types of persons, with control for item difficulty. There is no evidence to introduce any direct modification of the design of the tests. But it is important basic information on items which could be useful to test developers and to those who counsel candidates in effective test-taking strategies.

5. Basic research on the test-taking behavior of candidates. There is virtually no direct, empirical evidence on candidate behavior with regard to test-taking. Most of what is known is inferential, indirect, based upon statistical studies of item responses. Observational data collected at test administrations would be expensive and technically difficult (in the sense that it could not be disruptive, score-distorting, or violative of the candidates' rights as individuals. But observational data seems the only route to an understanding of some fundamental question about rate-of-work influences on score. Examples of such questions would be a) Does the Omit response truly require as much time as an overt mark? b) To what extent do candidates devote any time remaining to a review of the test? c) To what extent does the review of the test in the time remaining introduce score increments? d) Is there any evidence that minority candidates differ in their characteristic work patterns?

REFERENCES

- Bloom, Benjamin, and Broder, Lois J. Problem solving processes of college students. Supplementary Educational Monograph, 1950, No. 73.
- Brigham, Carl C. A Study of Error. New York: College Entrance Examinations Board, 1932.
- Connolly, John A., and Wantman, M.J. An exploration of oral reasoning processes in responding to objective test items. Journal of Educational Measurement, 1964, 1, 59-64.
- Cronbach, Lee J., and Warrington, W.G. "Time Limit Tests: Estimating Their Reliability and Degree of Speeding." Psychometrika, 1951, 14, 167-188.
- Gentile, J. Ronald. Toward an experimental analysis of reasoning on the Scholastic Aptitude Test: A pilot study. Princeton, N.J.: Educational Testing Service, 166 (RM-66-26).
- Gulliksen, Harold. Theory of Mental Tests. New York: John Wiley and Sons, 1950.
- Hoffman, Banesh. The Tyranny of Testing. New York: Crowell-Collier, 1962.
- Mueller, Danile J., and Wasser, Virginia. "Implications of Changing Answers on Objective Test Items." Journal of Educational Measurement, 1977, 14, 9-14.
- Stafford, Richard E. "The Speededness Quotient: A New Descriptive Statistic for Tests." Journal of Educational Measurement, 1971, 8, 275-278.
- Swineford, F. The test analysis manual. Princeton, New Jersey: Educational Testing Service, 1974 (SR-74-06).

APPENDIX A: Theoretical Values of $\bar{z}$ and σ_z for Selected Proportions Not Reaching
(Increments of .01 in Area Under the Normal Curve)

Proportion Not Reaching	$\bar{z}$	σ_z	Proportion Not Reaching	$\bar{z}$	σ_z
.01	-2.665217	0.311188	.26	-1.247559	0.496194
.01	-2.665217	0.333353	.27	-1.224612	0.500782
.03	-2.268066	0.348768	.28	-1.202226	0.505332
.04	-2.154346	0.361079	.29	-1.180359	0.509849
.05	-2.062714	0.371583	.30	-1.158975	0.514340
.06	-1.985385	0.380882	.31	-1.138043	0.518803
.07	-1.918116	0.389320	.32	-1.117534	0.523247
.08	-1.858330	0.397109	.33	-1.097420	0.527674
.09	-1.804341	0.404390	.34	-1.077677	0.532086
.10	-1.754984	0.411259	.35	-1.058283	0.536485
.11	-1.709414	0.417788	.36	-1.039215	0.540877
.12	-1.667005	0.424033	.37	-1.020456	0.545263
.13	-1.627270	0.430041	.38	-1.001988	0.549643
.14	-1.589834	0.435839	.39	-0.983793	0.554024
.15	-1.554393	0.441460	.40	-0.965856	0.558407
.16	-1.520699	0.446931	.41	-0.948163	0.562793
.17	-1.488550	0.452262	.42	-0.930700	0.567185
.18	-1.457778	0.457474	.43	-0.913454	0.571587
.19	-1.428238	0.462581	.44	-0.896413	0.575998
.20	-1.399809	0.467595	.45	-0.879566	0.580422
.21	-1.372387	0.472521	.46	-0.862903	0.584861
.22	-1.345880	0.477376	.47	-0.845412	0.589318
.23	-1.320209	0.482163	.48	-0.830085	0.593793
.24	-1.295305	0.486890	.49	-0.813912	0.598290
.25	-1.271106	0.491567	.50	-0.797885	0.602810

APPENDIX A: (Continued) Theoretical Values of $\bar{z}$ and σ_z for Selected Proportions Not Reaching (Increments of .01 in Area Under the Normal Curve)

Proportion Not Reaching	$\bar{z}$	σ_z	Proportion Not Reaching	$\bar{z}$	σ_z
.51	-0.781994	0.607356	.76	-0.409044	0.737411
.51	-0.766233	0.611930	.77	-0.294348	0.743725
.53	-0.750592	0.616535	.78	-0.379607	0.750179
.54	-0.735066	0.621171	.79	-0.364812	0.756783
.55	-0.719645	0.625843	.80	-0.349953	0.763548
.56	-0.704325	0.630552	.81	-0.335019	0.770487
.57	-0.689097	0.635301	.82	-0.320000	0.777614
.58	-0.673955	0.640092	.83	-0.304884	0.784943
.59	-0.658893	0.644928	.84	-0.289657	0.792494
.60	-0.643904	0.649812	.85	-0.274305	0.800287
.61	-0.628983	0.654747	.86	-0.258811	0.808344
.62	-0.614122	0.659736	.87	-0.243156	0.816693
.63	-0.599316	0.664782	.88	-0.227319	0.825366
.64	-0.584559	0.669889	.89	-0.211276	0.834402
.65	-0.569845	0.675059	.90	-0.194998	0.843846
.66	-0.555167	0.680279	.91	-0.178452	0.853754
.67	-0.540521	0.685606	.92	-0.161594	0.864197
.68	-0.525899	0.690991	.93	-0.144375	0.875265
.69	-0.511295	0.696456	.94	-0.126727	0.887079
.70	-0.496704	0.702007	.95	-0.108564	0.899801
.71	-0.482119	0.707647	.96	-0.089765	0.913670
.72	-0.467533	0.713382	.97	-0.070147	0.929058
.73	-0.452939	0.719221	.98	-0.049407	0.946620
.74	-0.438332	0.725167	.99	-0.026922	0.967804
.75	-0.423702	0.731228			